UNIVERSIDADE FEDERAL DE SERGIPE
CAMPUS ALBERTO CARVALHO
DEPARTAMENTO DE SISTEMAS DE INFORMAÇÃO

BRENO SANTANA SANTOS

ANÁLISE COMPARATIVA DE ALGORITMOS DE MINERAÇÃO DE TEXTO APLICADOS A HISTÓRICOS DE CONTAS PÚBLICAS

ITABAIANA
2015

UNIVERSIDADE FEDERAL DE SERGIPE
CAMPUS ALBERTO CARVALHO
DEPARTAMENTO DE SISTEMAS DE INFORMAÇÃO

BRENO SANTANA SANTOS

ANÁLISE COMPARATIVA DE ALGORITMOS DE MINERAÇÃO DE TEXTO APLICADOS A HISTÓRICOS DE CONTAS PÚBLICAS

Trabalho de Conclusão de Curso submetido ao Departamento de Sistemas de Informação da Universidade Federal de Sergipe como requisito parcial para a obtenção do título de Bacharel em Sistemas de Informação.

Orientador: Prof. Dr. Methanias Colaço Rodrigues Júnior

ITABAIANA
2015

Santana Santos, Breno.

Análise Comparativa de Algoritmos de Mineração de Texto Aplicados a Históricos de Contas Públicas / Breno Santana Santos – Itabaiana: UFS, 2015.
81f.

Trabalho de Conclusão de Curso em Bacharel em Sistemas de Informação – Universidade Federal de Sergipe, Curso de Sistemas de Informação, 2015.

1. Mineração de Texto. 2. Inteligência Artificial. 3. Sistemas de Informação. I. Análise Comparativa de Algoritmos de Mineração de Texto Aplicados a Históricos de Contas Públicas.

BRENO SANTANA SANTOS

ANÁLISE COMPARATIVA DE ALGORITMOS DE MINERAÇÃO DE TEXTO APLICADOS A HISTÓRICOS DE CONTAS PÚBLICAS

Trabalho de Conclusão de Curso submetido ao corpo docente do Departamento de Sistemas de Informação da Universidade Federal de Sergipe (DSIITA/UFS) como parte dos requisitos para obtenção do grau de Bacharel em Sistemas de Informação.

Itabaiana, 24 de Fevereiro de 2015.

BANCA EXAMINADORA:

_____

Prof(a) Methanias Colaço Rodrigues Júnior, Doutor.
Orientador
DSIITA/UFS

_____

Prof(a) Alcides Xavier Benicasa, Doutor
DSIITA/UFS

_____

Prof(a) André Vinícius Rodrigues Passos Nascimento, Mestre
DSIITA/UFS

A meus pais que me deram todo apoio, amor e compreensão para a realização de meus sonhos.

AGRADECIMENTOS

Oh my god! O chato do agradecimento é que temos pouco espaço para agradecer a todos que contribuíram. Farei um esforço para não esquecer daqueles que contribuíram para a conclusão desse trabalho. Desde já agradeço a todos que contribuíram de forma direta ou indireta para a realização deste trabalho.

Primeiramente, agradeço a Deus, pois sem Ele nada somos. Ele que sempre iluminou meu caminho, sempre me fez persistir e lutar pelo meu sonho.

Agradeço aos meus pais, meus exemplos de vida e porto seguro, Romualdo e Maria do Carmo, que sempre estiveram do meu lado, me dando apoio, conselhos, amor, carinho e umas surras para eu me orientar (algumas pareciam tentativas de homicídio kkk). Seus ensinamentos, princípios e valores foram fundamentais para eu alcançar meus objetivos sempre da melhor forma possível. Sou eternamente grato! Amo vocês!!!

Aos meus irmãos, Bruno e Brayon, que sempre me aturaram e me apoiaram quando mais precisei. Valeu meus Brothers! Amo vocês também!!!

A minha família pelos ensinamentos, preocupação, paciência, amor e momentos especiais, em especial aos meus avós paternos e maternos (in memoriam), aos meus tios e tias (né tia Carminha e tia Rosa?!), por fim, aos meus primos e primas (Roni, Serginho, Júnior Cabeção, Sandra e Vanessa! kkk).

A minha namorada, Alécia Alves, pelo amor, paciência e compreensão durante a minha ausência para a realização deste trabalho, assim como pela ajuda no estudo de caso com a seleção dos termos da área de saúde. Te amo, minha vida.

Aos meus amigos de longa data e pessoas especiais, bem como os que conheci durante a minha trajetória: Rony Peterson e família, Saulo Machado e família (Machado, o Pangalafumenga kkk), Luciana Melo, Savana e Josué Jr., Adriana e João (do zoío de buneca de feira kkk), Gilson e família, João Alves e família, Aline e família, Fábio da Coxinha e Luciene. Não poderia esquecer **a raça do Monteiro Lobato**, em especial: André Lucas (Xico Butico), Hemerson (Memé), Hugo Vinícius (O doido de Carira kkk), Carlos Eduardo (Dudu sou seu fã!!!), Diego Biribinha, Carlos Alberto, Vilker, Alan Balisa (Baiano doido da gota serena!!!), Roberto, Elias (fido canso!!! kkk) e Adison Chicleteiro (agora Chi Amedronta). Tem **a galera do IFS**, em especial: Leilane, Josivan, Roni (meu brother), Deyvisson, Luesia, Luiz Henrique (Super Aluno kkk), Thiago, Glauber, Kekel, Driele e Adriano. **Também aos amigos e pessoas especiais da UFS**, em especial: Fernanda, Nayra, Ythanna (quaterto fantástico :P), Igor

Peterson, Janisson Gois, Gilmar, Jéssica, Willams, Fabrício Barreto, Thiago, Nathan, Tauany, Kaline, Morgana, Cibele, Franciele, Maria Verônica e Clécia. Muito obrigado a todos pelo aprendizado e experiências, pelos momentos especiais, de alegria e de tristeza. Em resumo, obrigado por contribuírem em minha vida.

Ao meu orientador, Prof. Dr. Methanias Colaço R. Jr. (Methas Pai kkk) pela paciência e dedicação e pelos ensinamentos tanto profissionais quanto para a vida. Foi uma honra ser seu orientando e obrigado por tudo.

Aos amigos da Itatech, principalmente a Igor, Juli e Dósea, pela oportunidade que me foi dada para contribuir para a empresa.

Aos professores e técnico do DSI que sempre nos apoiaram e contribuíram para nosso aprendizado, em especial aos mestres e doutores amigos André Vinícius, Marcos Dósea e Alcides Benicasa, em que sou grato pelas orientações acadêmica e profissional, ensinamentos, conselhos, tanto para minha formação quanto para vida.

Muito obrigado a todos e sou fã de VOCÊS!!!

“Se não existe esforço, não existe progresso.”

(FREDERICH DOUGLASS)

SANTOS, Breno Santana. **Análise Comparativa de Algoritmos de Mineração de Texto Aplicados a Históricos de Contas Públicas**. 2015. Trabalho de Conclusão de Curso – Curso de Sistemas de Informação, Departamento de Sistemas de Informação, Universidade Federal de Sergipe, Itabaiana, 2015.

RESUMO

O uso de Mineração de Texto (MT) é importante para o processo de extração de conhecimento em bases textuais. Contudo, é importante avaliar se o conhecimento extraído ou gerado é relevante ou não para o usuário. Diante destas constatações, objetivou-se, com este trabalho, no âmbito das atividades de auditoria realizadas no Tribunal de Contas do Estado de Sergipe (TCE-SE), o desenvolvimento de um algoritmo de mineração de texto para a ferramenta TextMining (solução de MT do TCE-SE), bem como a avaliação de performance dos algoritmos de mineração de texto da ferramenta. Tal avaliação foi realizada mediante um estudo de caso nos históricos de contas públicas para detectar irregularidades no pagamento de diárias.

Palavras-chave: Algoritmos de Mineração de Texto. Histórico de Contas Públicas. Avaliação de Desempenho e Qualidade.

ABSTRACT

Using Text Mining (TM) is important in the process of knowledge extraction from text bases. However, it is important to assess if the knowledge extracted or produced is relevant or not to the user. Ahead of these verifications, it objectified, with this work, in the ambit of the audit activities performed in the Audit Office of the Country of Sergipe (AOC-SE), the development of a text mining algorithm for TextMining tool (MT solution of the AOC-SE), as well as performance evaluating of the tool text mining algorithms. This evaluation was performed by a case study in the public accounts of historical to detect irregularities in the payment of daily.

Key-words: *Text Mining Algorithms. Public Accounts History. Performance and Quality Assessment.*

LISTA DE FIGURAS

LISTA DE FÓRMULAS

LISTA DE GRÁFICOS

LISTA DE QUADROS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

DF	*Document Frequency*
DW	*Data Warehouse*
KDD	*Knowledge Discovery in Databases*
KDT	*Knowledge Discovery in Textual Databases*
MT	Mineração de Texto
PLN	Processamento de Linguagem Natural
RI	Recuperação da Informação
SAD	Sistemas de Apoio a Decisão
SGBD	Sistema de Gerenciamento de Banco de Dados
SISAP	Sistema de Auditoria Pública
TCE-SE	Tribunal de Contas do Estado de Sergipe
TF	*Term Frequency*
TFIDF	*Term Frequency – Inverse Document Frequency*
UFS	Universidade Federal de Sergipe

SUMÁRIO

1 INTRODUÇÃO

Na era da informação, esta passou a ser um dos maiores bens de uma organização, tendo o poder de influenciar no processo de tomada de decisão. Grandes massas de dados são geradas diariamente pelos sistemas que apoiam as atividades rotineiras das organizações, dificultando a tarefa analítica dos gestores. Diante dessa necessidade, surgiram os Sistemas de Apoio à Decisão (SADs) que, segundo Colaço Júnior (2004), permitem apoiar, contribuir e influenciar no processo de tomada de decisão. Os SADs permitem, a partir dos dados transacionais da organização, gerar informações gerenciais que facilitam o referido processo.

Como grande parte dos dados manipulados pelas organizações está em formato textual, torna-se fundamental o uso da técnica de Mineração de Texto (também conhecido por *Knowledge Discovery in Texts*, KDT, em inglês) para identificar padrões e conhecimentos para auxiliar nas decisões.

KDT é o processo de extração de informações, úteis e não-triviais, e conhecimento em texto desestruturado (VIJAYARANI; MUTHULAKSHMI, 2013). Para Magalhães (2008), o processo de Mineração de Texto é dividido em quatro etapas bem definidas: Seleção, Pré-processamento, Mineração e Assimilação.

Na *Seleção*, os documentos relevantes devem ser definidos para serem processados. No *Pré-processamento*, os documentos selecionados sofrerão um tratamento especial para que seja definida uma estrutura, a qual será utilizada na próxima etapa. Na *Mineração*, serão utilizadas técnicas para detectar os padrões não-visíveis nos dados. Por fim, na *Assimilação*, os usuários irão utilizar o conhecimento gerado para apoiar as suas decisões (BALINSKI, 2002; MAGALHÃES, 2008; SOUZA, 2011).

Para Wives (2002), o conhecimento gerado pode ser avaliado para determinar se o mesmo é relevante ou não para o usuário, ou seja, avaliar o desempenho do processo de mineração para a geração do conhecimento. Existem várias métricas, sendo as principais relacionadas ao desempenho, à acurácia, precisão e cobertura.

De forma análoga, o Tribunal de Contas de Sergipe (TCE-SE) lida com um imenso volume de informações, sendo necessária a utilização de mecanismos que tornem efetivas as atividades de auditoria.

Os tribunais de contas são instituições fundamentais para o processo de sustentação da democracia, agindo como regulamentador da aplicação dos recursos públicos (CASTRO, 2009). Em outras palavras, os tribunais de contas são órgãos fiscalizadores dos recursos públicos utilizados na Administração Pública, podendo responsabilizar os administradores pelos atos administrativos.

Auditoria é a atividade que realiza a validação das informações, verificação da obediência às normas e recomendações e avaliações dos controles em busca dos resultados da gestão (CASTRO, 2009).

Objetivando atender as necessidades do TCE-SE, o Departamento de Sistemas de Informação, do Campus Alberto Carvalho – UFS, desenvolveu uma aplicação que realiza a mineração de texto em qualquer campo descritivo de um sistema, a ferramenta TextMining.

A aplicação permite determinar se uma descrição é ou não evidência de irregularidade, tornando efetivo o trabalho do auditor na identificação de irregularidades. Para classificar uma descrição, a ferramenta dispõe de um algoritmo, Naïve Bayes, de forma parametrizada, especificando um limiar mínimo para auxiliar no processo classificatório. É importante destacar que existem três métodos para o Naïve Bayes: "Híbrido" (utilização da frequência do termo da amostra com *tf*, *term frequency*, da sentença), "Frequência Inversa" (*tfidf*, *term frequency – inverse document frequency*, da amostra com *tf* da sentença) e "Frequência" (frequência da amostra com frequência da sentença).

Este trabalho introduziu um segundo algoritmo, Similaridade, na ferramenta supracitada e foram avaliadas as métricas de qualidade e desempenho para as duas abordagens. A avaliação se deu por meio da coleta de métricas de tempo médio, acurácia, cobertura, medida F e precisão de cada algoritmo, bem como foi realizado um estudo de caso nos históricos de contas públicas custodiadas pelo TCE-SE, para analisar e comparar os resultados das métricas, conforme os objetivos descritos na seção 1.4.

1.1 Motivação

Pinho (2007) ressalta a necessidade do aperfeiçoamento do processo de obtenção de evidências com o auxílio da tecnologia. Assim, devido ao grande volume de dados e à dificuldade de realizar análise do conteúdo das prestações de contas, torna-se fundamental o uso de Mineração de Texto (MT) para extração de conhecimento de forma automática, com o

intuito de direcionar as atividades de auditoria (SOARES, 2010). Diante disso, faz-se necessário avaliar o conhecimento gerado pela técnica de MT para verificar se o mesmo é relevante ou não para apoiar as atividades de auditoria.

1.2 Justificativa

Este trabalho objetiva comparar o desempenho e qualidade de dois algoritmos de mineração de texto aplicados a históricos de contas públicas custodiadas pelo TCE-SE. A análise comparativa determinará o melhor algoritmo da ferramenta TextMining e, consequentemente, o conhecimento gerado por essa abordagem será efetivo e relevante para os auditores na descoberta de irregularidades como, por exemplo, a identificação de uma descrição de motivo de viagem a qual não é permitida o pagamento de diárias.

1.3 Objetivos do Trabalho

1.3.1 Objetivo Geral

Avaliar o desempenho e qualidade de algoritmos de mineração de texto aplicados a históricos de contas públicas custodiadas pelo Tribunal de Contas de Sergipe.

1.3.2 Objetivos Específicos

- Analisar a aplicação já desenvolvida e reaproveitar as rotinas de pré-processamento;
- Implementar um segundo algoritmo de mineração de texto, baseado em revisão bibliográfica sobre o uso do mesmo em campos descritivos;
- Definir como coletar as métricas de tempo médio, acurácia, precisão, medida F e cobertura de cada algoritmo de mineração;
- Realizar Estudo de Caso para analisar e comparar os resultados das métricas, determinando o melhor algoritmo com base no tempo médio, na acurácia, precisão, medida F e cobertura.

1.4 Organização da Monografia

O trabalho está organizado da seguinte forma.

No Capítulo 2, são apresentados a revisão bibliográfica e os conceitos necessários para a realização do trabalho. Inicia-se com o processo de descoberta de conhecimento em bases de dados e, em seguida, com o processo de mineração de texto, similaridade de documentos e avaliação de desempenho e qualidade, finalizando com os trabalhos relacionados.

No Capítulo 3, é apresentado um *overview* da solução para mineração de históricos. São abordadas informações e funcionalidades da ferramenta TextMining, algumas alterações e melhorias realizadas na aplicação e o algoritmo implementado.

No Capítulo 4, é apresentado o estudo de caso, objetivo, planejamento, seleção de participantes, dicionário utilizado e as métricas de avaliação de desempenho e qualidade utilizadas. Também é apresentada a execução do estudo de caso.

Já o Capítulo 5 conta com os resultados obtidos através do estudo de caso realizado e uma análise comparativa dos algoritmos.

Finalmente, no Capítulo 6, serão expostas as conclusões sobre o trabalho realizado e os possíveis trabalhos futuros relacionados.

2 REVISÃO BIBLIOGRÁFICA E CONCEITOS RELEVANTES AO TRABALHO

Este capítulo tem como objetivo explanar os conceitos necessários para o entendimento do trabalho, principalmente os conceitos relacionados ao processo de Descoberta de Conhecimento em Bases de Dados (KDD, *Knowledge Discovery in Databases*, em inglês), Mineração de Texto (KDT, *Knowledge Discovery in Texts*, em inglês), Similaridade de Documentos e Avaliação de Desempenho e Qualidade.

2.1 Descoberta de Conhecimento em Bases de Dados

KDD é o processo não-trivial de identificar padrões válidos, novos, potencialmente úteis em dados, ou seja, é o processo de descoberta de conhecimento ou padrões úteis e desconhecidos em grandes massas de dados (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996).

O processo de KDD consiste de vários passos, os quais envolvem preparação dos dados, busca por padrões, avaliação do conhecimento e refinamento, todos repetidos em múltiplas iterações (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996). KDD é composto por cinco passos bem definidos: Seleção, Pré-processamento, Transformação, Mineração de Dados, Análise / Assimilação, conforme é mostrado na Figura 01 (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996).

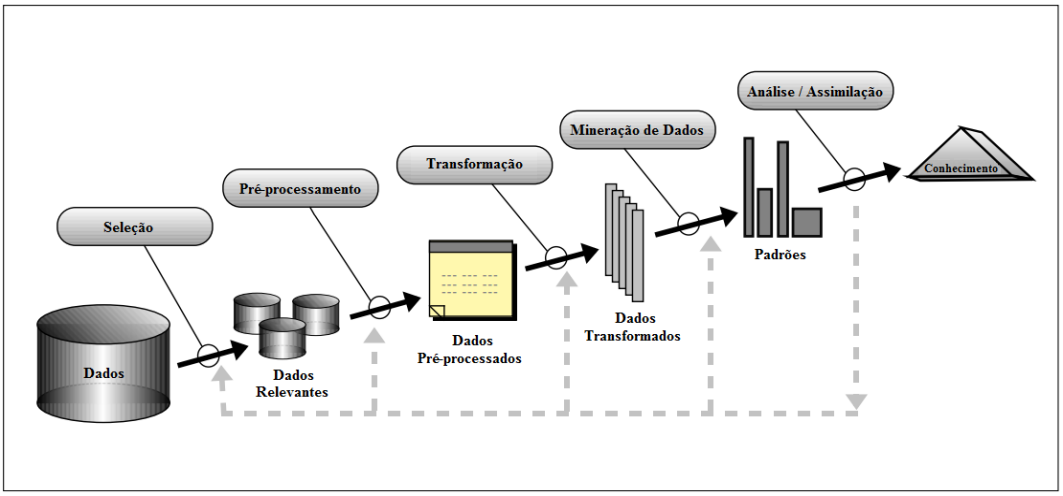

Figura 01. Passos que compõem o processo de KDD (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996).

Na etapa *Seleção*, serão definidas as fontes de dados relevantes, ou seja, as bases de dados importantes para o problema em questão, o qual se deseja resolver. No *Pré-processamento*, os dados serão tratados, pois como esses dados podem ser oriundos de diversas fontes, os mesmos podem conter divergência de valores e outras inconsistências. Na *Transformação*, os dados pré-processados serão convertidos para uma estrutura compatível com o algoritmo de mineração escolhido. Já na etapa *Mineração de Dados*, objetivo do processo de KDD, conforme Colaço Júnior (2004) complementa, é escolhida e executada uma técnica e algoritmo de mineração de acordo com o problema em questão, por exemplo, Classificação, Regressão, Agrupamento e Sumarização. E, por fim, na etapa de *Análise / Assimilação*, o conhecimento gerado será avaliado se é útil ou não para a tomada de decisão (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996).

Como é mostrado na Figura 01, o processo de KDD é um processo iterativo e interativo, em que o usuário participa e realiza decisões nas diversas etapas do processo, as quais podem também ser repetidas, dependendo do conhecimento gerado ou pela ausência do mesmo (COLAÇO JÚNIOR, 2004; FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996).

O processo de KDD pode ser aplicado em diversas áreas, incluindo Marketing, Finanças, Detecção de Fraudes, Manufaturas, Telecomunicações e Agentes da Internet. Conforme Colaço Júnior (2004) e Souza (2011), um exemplo clássico de utilização de KDD é o conhecimento descoberto nos dados da rede de supermercados Walmart. Foi descoberto que a maioria dos pais que iam comprar fraldas para seus filhos, acabavam comprando cerveja. Em uma jogada de marketing, as fraldas foram colocadas próximas da cerveja, sendo que as batatas-fritas estavam entre elas. Consequentemente, houve um aumento das vendas dos três produtos.

Outro exemplo de utilização do processo de KDD, segundo Bhandari *et al*. (1997), foi o uso do sistema ADVANCED SCOUT da IBM para ajudar os treinadores da NBA, no ano de 1996, a procurar e descobrir padrões interessantes nos dados dos jogos da NBA. Com esse conhecimento obtido, os treinadores podiam avaliar a eficácia das decisões de táticas e formular estratégias de jogo para jogos futuros. O sistema foi distribuído para dezesseis das vinte e nove equipes da NBA, sendo usado de forma efetiva por algumas equipes para a preparação de jogadas e processos analíticos, como foi o caso do time Seattle Supersonics, o qual atingiu as finais da NBA.

2.2 Mineração de Texto

Depois de entendido o processo de KDD, torna-se necessária a explicação do processo de KDT ou Mineração de Texto, principal conceito para o entendimento deste trabalho.

Mineração de Texto ou KDT é o processo de descoberta de conhecimento, potencialmente útil e previamente desconhecimento, em bases de dados desestruturadas, ou seja, extração de conhecimento útil para o usuário em bases textuais (BALINSKI, 2002; FELDMAN; DAGAN, 1995; MAGALHÃES, 2008; SOUZA, 2011).

Para Magalhães (2008), o processo de Mineração de Texto é dividido em quatro etapas bem definidas: Seleção, Pré-processamento, Mineração e Assimilação, conforme é mostrado na Figura 02.

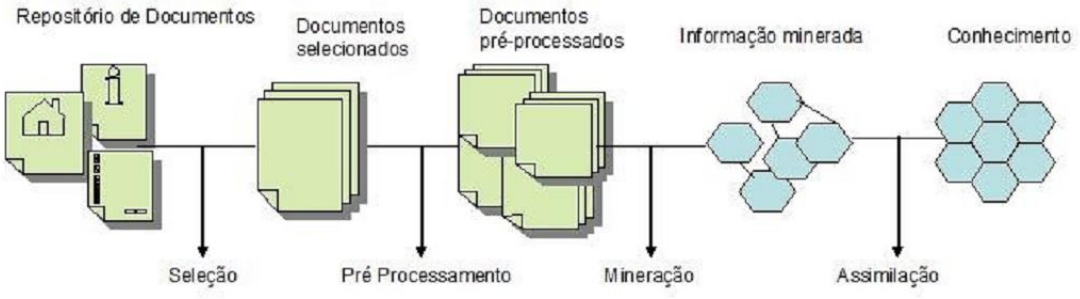

Figura 02. Processo de Mineração de Texto (MAGALHÃES, 2008).

Conforme Balinski (2002), Magalhães (2008) e Souza (2011) afirmam, na *Seleção*, os documentos relevantes devem ser escolhidos, os quais serão processados. No *Pré-processamento*, ocorrerá a conversão dos documentos em uma estrutura compatível com o minerador, bem como ocorrerá um tratamento especial do texto. Na *Mineração*, o minerador irá detectar os padrões com base no algoritmo escolhido. E por fim, na *Assimilação*, os usuários irão utilizar o conhecimento gerado para apoiar as suas decisões.

É notório a semelhança entre os processos de KDD e KDT, sendo que no KDT não possui a etapa de *Transformação*. O fato da ausência da etapa *Transformação*, etapa no processo de KDD que converte os dados pré-processados para uma estrutura utilizada na etapa de *Mineração de Dados*, é que a etapa de *Pré-processamento* no KDT além de realizar um tratamento no texto, permite definir uma estrutura compatível com as entradas dos algoritmos de mineração.

Para Magalhães (2008) e Souza (2011), a etapa *Pré-processamento* pode ser dividida em quatro subetapas: Remoção de *StopWords*, Conflação, Normalização de Sinônimos e Indexação.

Na etapa *Remoção de StopWords*, os termos com pouca ou nenhuma relevância para o documento serão removidos (SOUZA, 2011). São palavras auxiliares ou conectivas, ou seja, não são discriminantes para o conteúdo do documento (MAGALHÃES, 2008; WIVES, 2002). São, em sua maioria, pronomes, preposições, artigos, numerais e conjunções (SÁ, 2008; SOARES, 2010). Para auxiliar na remoção das *stopwords*, geralmente, utiliza-se uma lista de *stopwords* (BRAMER, 2007; SÁ, 2008). Para facilitar o entendimento, na Figura 03 é apresentado um exemplo de remoção de *stopwords*.

Figura 03. Exemplo de Remoção de *StopWords* (SOUZA, 2011).

Na etapa seguinte, *Conflação*, realiza-se uma normalização morfológica, ou seja, realiza-se uma combinação das palavras que são variantes morfológicas em uma única forma de representação. Um dos procedimentos mais conhecidos de conflação é Radicalização (*Stemming*) (GONZALEZ; LIMA, 2003).

Na radicalização, as palavras são reduzidas ao seu radical, ou seja, as palavras variantes morfologicamente serão combinadas em uma única representação, o radical. (BRAMER, 2007; MAGALHÃES, 2008; WEISS; INDURKHYA; ZHANG, 2010). Para Magalhães (2008), a radicalização pode ser efetuada com o auxílio de algoritmos de radicalização, sendo os mais utilizados o algoritmo de Porter (*Porter Stemming Algorithm*) e algoritmo de Orengo (*Stemmer Portuguese* ou RLSP). A Figura 04 exemplifica o processo de radicalização de um texto utilizando o algoritmo de Porter.

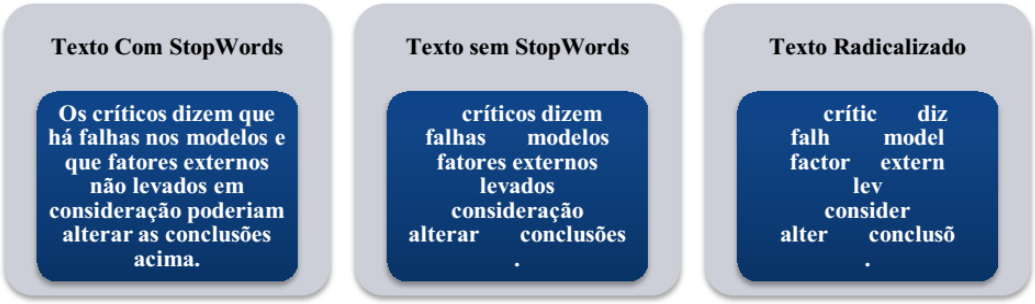

Figura 04. Exemplo de Radicalização utilizando o algoritmo de Porter (SOUZA, 2011).

Em seu estudo, Orengo (2001) identificou dois problemas no processo de radicalização:

- *Overstemming*: quando a *string* removida não é um sufixo, mas sim parte do radical da palavra. Isso possibilita a combinação de palavras não relacionadas;
- *Understemming*: quando parte do sufixo não é removido, ocasionando numa falha de conflação de palavras relacionadas.

Após a conflação, na etapa *Normalização de Sinônimos*, os termos que possuem significados similares serão agrupados em um único termo, por exemplo, as palavras *ruído*, *tumulto* e *barulho* serão substituídas ou representadas pelo termo *barulho* (BALINSKI, 2002; MAGALHÃES, 2008; SOUZA, 2011).

Na normalização de sinônimos, é formado um vocabulário controlado que, segundo Wives (2002), é a utilização de termos adequados para representar um documento, sendo esses termos pré-definidos e específicos a um determinado assunto específico de uma área. Isso facilita a busca, pois os termos são comumente utilizados pelos usuários da área.

E, por fim, na etapa *Indexação*, atribui-se uma pontuação para cada termo, garantindo uma única instância do termo no documento (SOUZA, 2011). Para Balinsky (2002), Magalhães (2008) e Wives (2002), no processo de atribuição de pesos devem ser considerados dois pontos: (a) quanto mais vezes um termo aparece no documento, mais relevante ele é para o documento; (b) quanto mais vezes um termo aparece na coleção de documentos, menos importante ele é para diferenciar os documentos.

Existem várias formas de determinar o peso de um termo (pontuação), conforme Balinsky (2002), Bramer (2007), Magalhães (2008) e Wives (2002), os principais métodos de pontuação são:

- Booleano ou Binário: o peso para um determinado termo será 1 se o mesmo aparece no documento. Caso contrário, o peso será 0. Indica a presença ou ausência do termo no documento.
- Frequência do Termo (*term frequency* ou *tf*): o peso é a frequência do termo no documento. Consiste da razão entre a quantidade de vezes que o termo apareceu no documento e a quantidade total de termos contidos no documento, como é mostrado na Fórmula 01, onde n_i é a quantidade de ocorrências do termo i no documento e $|D|$ a quantidade total de termos no documento.

$$tf(termo\,i) = \frac{n_i}{|D|}$$

Fórmula 01. Fórmula para calcular a frequência do termo.

- Frequência do Documento (*Document Frequency* ou *df*): é o número de documentos que possui um determinado termo.
- Frequência Inversa do Documento (*Inverse Document Frequency* ou *idf*): refere-se à importância de um termo em um conjunto de documentos. Quanto maior o *idf*, mais representativo é o termo para o documento. Consiste no logaritmo da razão entre o número total de documentos e a frequência do documento, conforme é demonstrado na Fórmula 02, onde $|N|$ é a quantidade total de documentos e *df(termo i)* a frequência do documento para o termo i.

$$idf(termo\,i) = log\,\frac{|N|}{df(termo\,i)}$$

Fórmula 02. Fórmula para calcular a frequência inversa do termo.

- *tfidf* (*Term Frequency – Inverse Document Frequency*): o peso para o termo é associado na proporção da frequência do termo no documento e na proporção inversa do número de documentos na coleção em que o termo aparece pelo menos uma vez, ou seja, combina o *tf* com *idf*, como é

mostrado na Fórmula 03, onde *tf(termo i)* e *idf(termo i)* são, respectivamente, o *tf* e *idf* do termo i. Obtém-se, assim, o índice de maior representatividade do termo.

$$tfidf(termo\,i) = tf(termo\,i) \times idf(termo\,i)$$

Fórmula 03. Fórmula para calcular o *tfidf* do termo.

As subetapas do *Pré-processamento* permitem uma redução da dimensionalidade do texto, pois, de acordo com Balinsky (2002), Magalhães (2008) e Souza (2011), um documento pode ser representado por um vetor de termos. Como um termo representa uma dimensão do texto, quanto maior a dimensionalidade do texto, mais complexa será a análise feita pelo algoritmo de mineração.

Como no *Pré-processamento* definimos uma estrutura para os dados desestruturados, de acordo com Feldman e Dagan (1995), devido às limitações severas da tecnologia atual de processamento robusto de texto, devemos optar por estrutura bastante simples, que permita automaticamente a extração de texto e a um custo razoável.

Assim como no KDD, o processo de Mineração de Texto possui diversas aplicações como, por exemplo, extração de palavras-chave, determinação de sistemas representacionais preferenciais, classificação de documentos por categoria, filtro de documentos e entre outras.

2.3 Similaridade de Documentos

Como foi dito na seção anterior, um documento pode ser considerado um vetor de termos. Balinsky (2002), Magalhães (2008) e Wives (2002) afirmam que cada elemento do vetor é considerado uma coordenada dimensional e os documentos podem ser colocados num espaço euclidiano de *n* dimensões (*n* é o número de termos). A posição do documento em cada dimensão é dada pelo peso (pontuação calculada na fase de Indexação).

A distância entre um documento e outro é o grau de similaridade (WIVES, 2002). Documentos que possuem os mesmos termos acabam sendo colocados numa mesma região no espaço euclidiano, ou seja, são similares. Para Wives (2002), os vetores podem ser comparados e o grau de similaridade pode ser identificado.

Weiss; Indurkhya e Zhang (2010) e Wives (2002) afirmam que a similaridade entre dois documentos pode ser obtida pelos termos que ocorrem em ambos, ou seja, pelos termos compartilhados. Os documentos mais similares são os que possuem mais termos em comum.

Ainda segundo Weiss; Indurkhya e Zhang (2010), no cálculo da similaridade, são ignorados os termos que ocorrem em um documento e que não ocorrem no outro. Em outras palavras, só interessam os termos que ocorrem nos dois, isto é, a ocorrência positiva desse em ambos.

Similaridade é considerada o coração do método de classificação *K-Nearest-Neighbor*. A diferença entre ambos é que no *K-Nearest-Neighbor* consideram-se os *k* documentos mais similares, a depender do valor de *k*, podem ser considerados os documentos com *score* inferior aos de maior *score* para determinar a classe do novo documento (WEISS; INDURKHYA; ZHANG, 2010).

Conforme Weiss; Indurkhya e Zhang (2010), Similaridade considera apenas os documentos com maior *score* e a classe do novo documento será a classe que mais ocorre nesses. É importante frisar que para o cálculo do grau de similaridade (*score*), devem ser apenas considerados os termos em comum.

Magalhães (2008), Weiss; Indurkhya e Zhang (2010) e Wives (2002) afirmam que existem várias formas de calcular o grau de similaridade, isto é, as funções de similaridade. Depois de calcular os *scores*, podemos criar uma lista em forma de *ranking*, em que os documentos mais similares estão no topo da lista.

De acordo com Magalhães (2008), Souza e Claro (2014), Weiss; Indurkhya e Zhang (2010) e Wives (2002), as principais funções de similaridade são:

- Contagem de Palavras: é considerada a função mais simples de mensurar a similaridade, pois se baseia apenas na contagem de termos que ocorrem em ambos documentos, isto é, as ocorrências positivas dos termos.
- Contagem de Palavras com Bônus: De forma análoga à contagem de palavras, serão contabilizados os termos em comum aos vetores com apenas um diferencial, para cada termo analisado, se esse termo ocorre em ambos documentos, será adicionado um bônus ao *score*, conforme é visto na Fórmula 04, onde *K* é a quantidade total de termos do novo documento, *w(j)* a pontuação para o termo j, *D(i)* o documento i da coleção e a expressão *1/df(j)* o bônus para o termo j. O bônus é considerado uma variação do *idf*.

Se o termo ocorre em muitos documentos, o valor do bônus é baixo. Já se o termo aparece em poucos, o bônus é alto.

$$Similarity(D(i)) = \sum_{j=1}^{K} w(j),$$
$$w(j) = \begin{cases} 1 + 1/df(j), & se\ o\ termo\ j\ ocorre\ em\ ambos\ documentos \\ 0, & caso\ contrário \end{cases}$$

Fórmula 04. Fórmula para a Contagem de Palavras com Bônus.

- *Cosine Similarity*: função de similaridade mais utilizada no campo de Recuperação de Informação (RI) para comparar documentos. Representa o cosseno do ângulo formado por dois vetores, como é mostrado na Fórmula 05, onde d_1 e d_2 são os documentos cuja similaridade será calculada, $w_{d1}(j)$ o peso do termo j em d_1, $w_{d2}(j)$ o peso do termo j em d_2, $\sqrt{\sum (w_{d1}(j))^2}$ a normalização de d_1 e $\sqrt{\sum (w_{d2}(j))^2}$ a normalização de d_2. Quanto mais próximo de zero for o valor do cosseno, menos similares são os documentos. Já quando for mais próximo de um, mais similares são.

$$cos(d_1, d_2) = \frac{\sum (w_{d1}(j) \times w_{d2}(j))}{\sqrt{\sum (w_{d1}(j))^2} \times \sqrt{\sum (w_{d2}(j))^2}}$$

Fórmula 05. Fórmula para a *Cosine Similarity*.

- Distância Euclidiana: representa a menor distância entre dois vetores de termos no espaço euclidiano, como é visto na Fórmula 06, em que d_1 e d_2 são os documentos, K o número de termos, $w_{d1}(j)$ o peso do termo j em d_1 e $w_{d2}(j)$ o peso do termo j em d_2.

$$dist(d_1, d_2) = \sqrt{\sum_{j=1}^{K} (w_{d1}(j) - w_{d2}(j))^2}$$

Fórmula 06. Fórmula para a Distância Euclidiana.

- Distância de *Manhattan*: é a soma das distâncias absolutas sem cada dimensão. Corresponde à distância a ser percorrida para chegar de um ponto a outro, em que o caminho é percorrido em quadras, conforme é mostrado na Fórmula 07, onde d_1 e d_2 são os documentos, K o número de termos, $w_{d1}(j)$ o peso do termo j em d_1 e $w_{d2}(j)$ o peso do termo j em d_2.

$$dist(d_1, d_2) = \sum_{j=1}^{K} |w_{d1}(j) - w_{d2}(j)|$$

Fórmula 07. Fórmula para a Distância *Manhattan*.

- Produto Escalar: corresponde ao somatório do produto dos pesos de um termo em dois documentos, como é visto na Fórmula 08, onde d_1 e d_2 são os documentos, K o número de termos, $w_{d1}(j)$ o peso do termo j em d_1 e $w_{d2}(j)$ o peso do termo j em d_2.

$$sim(d_1, d_2) = \sum_{j=1}^{K} w_{d1}(j) \times w_{d2}(j)$$

Fórmula 08. Fórmula para o Produto Escalar.

Weiss; Indurkhya e Zhang (2010) afirmam que o novo documento será comparado com todos da coleção para determinar o grau de similaridade entre o novo documento e cada documento da coleção. Comparar sequencialmente os novos documentos com todos da coleção é um processo ineficiente (WEISS; INDURKHYA; ZHANG, 2010).

Para facilitar o entendimento sobre Similaridade, a Figura 05 demonstra o cálculo da similaridade entre um novo documento e todos documentos do dicionário, utilizando a função de similaridade *Contagem de Palavras*. Como podemos ver, foi calculado o *score* entre o novo documento e todos do dicionário por meio da contagem de palavras cuja ocorrência em ambos foi positiva, isto é, a contabilização delas que ocorrem em ambos, ignorando as que ocorrem apenas em um e as ausentes em ambos. Existem dois documentos que possuem o maior *score*, grau de similaridade igual a 2. Como os dois documentos com maior score possuem classe igual a um (última coluna do dicionário), a classe do novo documento também será um.

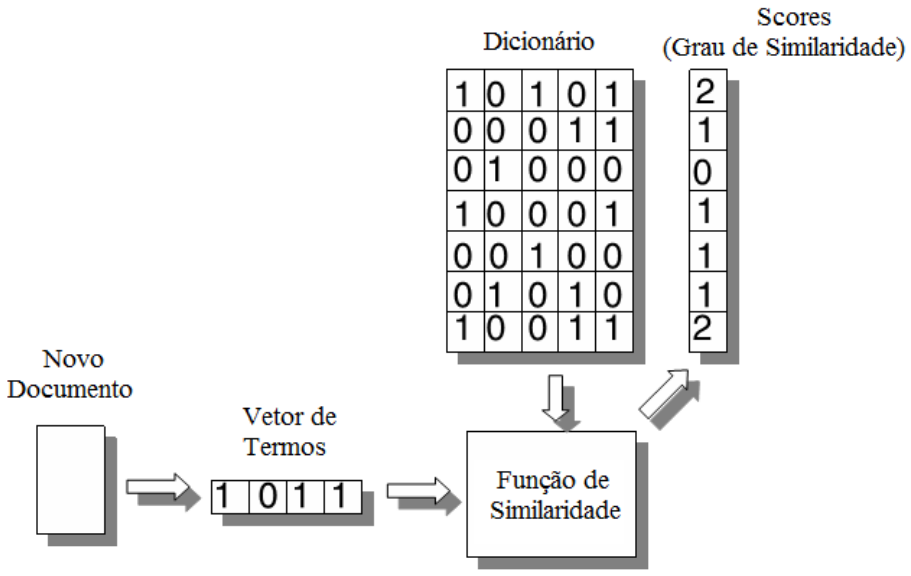

Figura 05. Cálculo de Similaridade dos Documentos (WEISS; INDURKHYA; ZHANG, 2010).

2.4 Avaliação de Desempenho e Qualidade

Existem diversas formas de avaliar a capacidade de predição de um classificador para determinar a classe de vários registros (HAN; KAMBER; PEI, 2011). Segundo Han; Kamber e Pei (2011) e Witten e Frank (2005), a matriz de confusão é a forma mais simples de analisar o desempenho e qualidade de um classificador em reconhecer registros de diferentes classes.

Em outras palavras, matriz de confusão é um recurso que permite demonstrar o

desempenho de um classificador, ou seja, a frequência com que os registros de classe X foram corretamente classificados como classe X ou, até mesmo, classificados erroneamente como outra classe (BRAMER, 2007).

De acordo com Bramer (2007) e Han; Kamber e Pei (2011), para *n* classes, a matriz de confusão é uma tabela de dimensão *n* x *n*. Para cada classificação possível existe uma linha e coluna correspondente, ou seja, os valores das classificações serão distribuídos na matriz de confusão de acordo com os resultados, assim gerando a matriz de confusão para as classificações realizadas. Ainda conforme Bramer (2007), as linhas correspondem às classificações corretas e as colunas representam as classificações realizadas pelo classificador. Por exemplo, na Tabela 01, o valor $V_{1,1}$ corresponde ao número de registros de classe 1 em que foram classificados com classe 1 pelo classificador.

Tabela 01. Matriz de Confusão para *n* classes.

Classe Atual	**Classificado como**			
	Classe 1	Classe 2	...	Classe *n*
Classe 1	$V_{1,1}$	$V_{1,2}$	...	$V_{1,n}$
Classe 2	$V_{2,1}$	$V_{2,2}$	...	$V_{2,n}$
...	...	...	...	...
Classe *n*	$V_{n,1}$	$V_{n,2}$	...	$V_{n,n}$

Quando existem apenas duas classes, uma é considerada como "positive" e a outra como "negative" (BRAMER, 2007). Para Bramer (2007), Han; Kamber e Pei (2011) e Witten e Frank (2005), os valores da matriz de confusão são referenciados como *true* e *false positives* e *true* e *false negatives*, como é visto na Tabela 02.

Tabela 02. Matriz de Confusão para duas classes.

Actual class	**Predicted class**	
	Positive	**Negative**
Positive	*True Positive (TP)*	*False Negative (FN)*
Negative	*False Positive (FP)*	*True Negative (TN)*

Conforme Bramer (2007), Han; Kamber e Pei (2011) e Witten e Frank (2005), bem como pode ser visto na Tabela 02, existem quatro situações:

- *True Positive (TP)*: é o número de instâncias de classe *positive* que foram classificadas como *positive*;
- *False Positive (FP)*: é o número de instâncias de classe *negative* que foram

classificadas como *positive*;

- *False Negative (FN)*: é o número de instâncias de classe *positive* que foram classificadas como *negative*;
- *True Negative (TN)*: é o número de instâncias de classe *negative* que foram classificadas como *negative*.

Bramer (2007), Han; Kamber e Pei (2011) e Witten e Frank (2005) afirmam que a avaliação de um classificador se dará pela análise dos valores nela contidos, bem como na verificação do somatório dos elementos das diagonais principal e secundária. Um bom classificador é aquele que possui a soma da diagonal principal maior que a da secundária. Um classificador é considerado ideal quando a soma da diagonal secundária é igual a zero, contudo esse será considerado um péssimo classificador se possuir o somatório da diagonal principal igual a zero.

De posse dos valores da matriz de confusão, podem ser utilizadas as métricas de avaliação de desempenho e qualidade de um classificador (WITTEN; FRANK, 2005). As principais métricas de desempenho e qualidade para Bramer (2007), Han; Kamber e Pei (2011) e Witten e Frank (2005) são:

- Acurácia (*accuracy*): é o percentual de instâncias classificadas corretamente, como é mostrado na Fórmula 09.

$$acurácia = \frac{TP + TN}{TP + TN + FP + FN}$$

Fórmula 09. Fórmula para o cálculo da Acurácia.

- Precisão (*precision*): é o percentual de instâncias classificadas como *positive* que são realmente *positive* (Fórmula 10).

$$precisão = \frac{TP}{TP + FP}$$

Fórmula 10. Fórmula para o cálculo da Precisão.

- Cobertura ou Revocação (*recall*): é o percentual de instâncias *positive* que foram classificadas corretamente como *positive* (Fórmula 11).

$$revocação = \frac{TP}{TP + FN}$$

Fórmula 11. Fórmula para o cálculo da Revocação.

- Medida F (*F1 Score*): é a medida que combina a precisão e revocação (cobertura), ou seja, é a média harmônica da precisão e revocação (Fórmula 12).

$$MedidaF = \frac{2 \times precisão \times revocação}{precisão + revocação}$$

Fórmula 12. Fórmula para o cálculo da Medida F.

2.5 Trabalhos Relacionados

Diante dos diversos estudos encontrados, que utilizam técnicas de Mineração de Texto de forma direta ou indireta, destacam-se os trabalhos realizados por Souza (2011), Balinski (2002) e Magalhães (2008).

No trabalho realizado por Souza (2011), o mesmo concebeu uma aplicação para descoberta de características psicológicas dos indivíduos. Com base nos e-mails dos participantes da lista de discussão SergInfo, a ferramenta, por meio de Mineração de Texto, determina o Sistema Representacional Predominante para aquele contexto.

No segundo, o autor realiza um filtro (sistema de filtragem) num software de correio eletrônico jurídico, Direto, utilizando um padrão de linguagens de filtro de mensagens, Sieve. Propôs também um serviço de canais de comunicação, utilizando técnicas de classificação de texto, para a divulgação das informações necessárias nesses canais de comunicação, conforme as necessidades dos usuários. O filtro possibilitou um aumento da produtividade dos usuários.

No terceiro trabalho, a autora desenvolveu um classificador de documentos jurídicos que permite buscar casos julgados similares a um outro descrito pelo usuário (seja um documento submetido ou texto livre informado) em uma base jurisprudencial. A ferramenta

auxilia os profissionais do Direito na análise de um processo e, consequentemente, em sua fundamentação jurídica.

No contexto da utilização de Mineração de Texto nas atividades de auditoria em históricos de contas públicas, destaca-se o trabalho de Soares (2010). O autor definiu um processo de MT com o intuito de classificar despesas públicas por objeto de gasto, por meio dos campos de notas de empenho nos históricos de contas públicas sob custódia do Tribunal de Contas dos Municípios do Estado do Ceará (TCMCE). Para implementar a solução, o mesmo utilizou o modelo de projeto CRISP-DM, o SGBD PostgreSQL e a ferramenta Weka, bem como utilizou os dados armazenados no SIM (Sistema de Informações Municipais) para realizar a mineração nos documentos de nota de empenho.

No âmbito de pesquisas relacionadas à comparação de algoritmos de mineração de dados, destacam-se os trabalhos de Amooee; Minaei-Bidgoli e Bagheri-Dehnavi (2011), Ghazvini; Awwalu e Bakar (2014) e Brilhadori e Lauretto (2013).

No primeiro trabalho, os autores realizaram um processo de mineração de dados na indústria Ahanpishegan para prever futuras falhas em peças, com base nos históricos de ocorrências de peças defeituosas. Para detectar peças defeituosas, foram utilizados diferentes algoritmos (árvores de decisão CHAID, C&R e QUEST, Redes Neurais, Redes Bayesianas, logistic regression e SVM), assim como foi realizada uma análise comparativa da acurácia e do tempo de processamento desses algoritmos. Após a análise, foi constatado que o algoritmo SVM obteve o melhor tempo de processamento e acurácia global. Já os algoritmos C&R e QUEST tiveram os piores tempo de processamento, mas obtiveram as melhores porcentagens de acurácia. Por fim, o algoritmo de Redes Neurais obteve a pior acurácia.

No segundo, os autores compararam dois algoritmos de mineração de dados, Naïve Bayes e Multilayer Perceptron, para classificar notas, como sendo falsas ou reais, de um banco de notas da *University of California Irvine* (UCI) sobre dois pontos de vista: *holdout* e *cross validation*. Após a realização dos experimentos, concluiu-se que o Multilayer Perceptron obteve melhores resultados do que o Naïve Bayes em termos de acurácia em ambos pontos de vista (*holdout* e *cross validation*). Contudo, vale ressaltar que o Naïve Bayes foi o mais rápido (melhor em tempo de classificação) do que o Multilayer Perceptron.

No terceiro, os autores realizaram uma análise comparativa de desempenho de um algoritmo de árvores de classificação (J48) e um algoritmo de máquinas de suporte vetorial (SMO) quando combinados na forma de *ensembles bagging* e *boosting* (*AdaBoost.M1*), todos

implementados dentro do ambiente Weka. A análise foi baseada em experimentos de validação cruzada sobre 21 conjuntos de dados disponíveis no *UCI Machine Learning Repository*, sendo que 8 possuíam classes binárias e 13 apresentavam multiclasses. Os resultados demonstraram que o J48 possui maiores acurácias sob a forma de *ensembles*, principalmente na configuração *Boosting*. Já o SMO parece menos sensível às configurações *ensembles* utilizadas, ou seja, não foi encontrado indício de que o SMO tenha sua performance influenciada pelas configurações *ensemble*. Vale ressaltar que ambos os algoritmos obtiveram números de vitórias similares entre os conjuntos de dados.

Por fim, no contexto de pesquisas relacionadas à comparação de algoritmos de mineração de texto, destacam-se os trabalhos de Lamkanfi *et al.* (2011), Ting; Ip e Tsang (2011), Vijayarani e Muthulakshmi (2013) e Mccallum e Nigam (1998).

No primeiro trabalho, os autores realizaram um processo de mineração de texto para classificar erros reportados em dois projetos *open source*, Eclipse e GNOME, por meio do Bugzilla, sistema de rastreamento de erros. Para prever o tipo de gravidade do erro, foram utilizados diferentes algoritmos do ambiente Weka (Naïve Bayes, Naïve Bayes Multinomial, K-Nearest Neighbor e Support Vector Machines), assim como foi realizada uma análise comparativa em relação à acurácia e tamanho base de treinamento. Após a análise, foi constatado que o Naïve Bayes Multinomial obteve melhor performance em relação aos outros algoritmos.

No segundo, os autores realizam uma análise comparativa dos algoritmos Naïve Bayes, SVM (SMO), K-Nearest Neighbour (*lazy* IBk), árvore de decisão J48, todos disponíveis na ferramenta Weka. Para tal análise, foi utilizado um conjunto de dados de 4000 documentos classificados em quatro diferentes classes: *business*, *politic*, *sports*, e *travel*. O conjunto de treinamento era constituído de 1200 documentos (30% do total de documentos), já o de teste era composto pelos 2800 documentos restantes. Ao final da análise, foi comprovado que o Naïve Bayes era o melhor algoritmo em termos de acurácia, precisão, cobertura e medida F.

No terceiro, as autoras analisaram o desempenho dos classificadores bayesianos e *lazy* para classificar arquivos que estão armazenados em um disco rígido de um computador. Foram escolhidos cinco algoritmos, sendo dois classificadores bayesianos, BayesNet e Naïve Bayes, e três classificadores *lazy*, IBL (*Instance Based Learning*), IBK (*K-Nearest Neighbour*) e Kstar, todos disponíveis na ferramenta Weka. Esta foi utilizada para analisar o desempenho dos algoritmos em um conjunto de dados, o qual possui 80000 instâncias e quatros atributos

(nome, tamanho, extensão e caminho do arquivo). Inicialmente foram analisados os desempenhos de BayesNet e Naïve Bayes, sendo que o primeiro obteve os melhores resultados. Da mesma forma, os classificadores *lazy* foram avaliados. Foi constatado que o algoritmo IBK foi a melhor abordagem *lazy*. Por fim, foi realizada a análise comparativa entre BayesNet e IBK. Após a verificação dos resultados, os classificadores *lazy* são mais eficientes do que os bayesianos, sendo o IBK a melhor técnica dentre as demais analisadas.

Por fim, no quarto trabalho, os autores explanaram conceitos relacionados a dois modelos de classificadores bayesianos, multi-variate Bernoulli (rede bayesiana que considera apenas a presença e ausência dos termos) e Naïve Bayes Multinomial (considera a frequência dos termos), bem como realizaram uma análise comparativa dessas abordagens em cinco conjuntos de dados. Os conjuntos de dados foram: **Yahoo**, páginas web apontadas pelo Yahoo Science (95 classes); **Industry Sector**, páginas web de companhias classificadas por setor industrial (71 classes); **Newsgroups**, artigos uniformemente divididos entre grupos de discussão UseNet (20 classes); **WebKB**, páginas web obtidas em departamentos de ciência da computação (4 classes); **Reuters**, parte do conjunto de dados Reuters-21578, "ModApte", que contém artigos *newswire* da agência Reuters (10 classes). Após a verificação dos resultados, foi constatado que o modelo multi-variate Bernoulli, às vezes, possui melhor desempenho do que o multinomial em vocabulários de tamanho pequeno. Contudo, geralmente, o multinomial supera o multi-variate Bernoulli em vocabulários de tamanho grande e possui, em média, uma redução de 27% na taxa de erro comparado ao multi-variate Bernoulli.

3 OVERVIEW DA SOLUÇÃO PARA MINERAÇÃO DE HISTÓRICOS

Neste capítulo, serão apresentadas as principais informações sobre a ferramenta TextMining, bem como as alterações e melhorias efetuadas na mesma e o algoritmo implementado. Na seção 3.1, serão abordados assuntos referentes às funcionalidades da ferramenta TextMining. Em seguida, na 3.2, as alterações e melhorias efetuadas na aplicação e, por fim, na 3.3, apresentação do algoritmo implementado.

3.1 Ferramenta TextMining

O Departamento de Sistemas de Informação, do Campus Prof. Alberto Carvalho – UFS, de posse de uma cópia do DW[1] do sistema SISAP[2], a qual foi cedida pelo TCE-SE, desenvolveu uma aplicação que realiza a mineração de texto em qualquer campo descritivo de um sistema.

A aplicação permite determinar se as informações são ou não evidências de irregularidades, ou seja, se uma descrição está ou não de acordo com a lei e com o que se espera dos jurisdicionados. Desta forma, a ferramenta tem como objetivo tornar efetivo o trabalho do auditor na identificação de irregularidades. Suas principais funcionalidades são os gerenciamentos de perfis, de dicionários e de classificações. Considera-se gerenciamento o conjunto de funções relacionadas ao cadastro, edição, consulta, exclusão e visualização de informações.

Iniciando pelo gerenciamento de perfis, estes são mecanismos que auxiliam nas consultas por meio dos filtros anexados aos perfis. Conforme é mostrado na Figura 06, é por meio deles que o usuário poderá determinar dinamicamente os campos que deseja filtrar nas telas, nas quais poderá escolher o perfil. Na TextMining, está disponível para o usuário as funcionalidades de cadastro, consulta e exclusão.

[1] DW: do inglês *Data Warehouse* (Armazém de Dados), corresponde a um banco de dados histórico que auxilia o processo de tomada de decisão (COLAÇO JÚNIOR, 2004).
[2] SISAP: Sistema de Auditoria Pública, um banco de dados com informações orçamentárias, financeiras, contábeis e administrativas dos órgãos sob jurisdição do TCE-SE (http://www.tce.se.gov.br/sitev2/sisap.php).

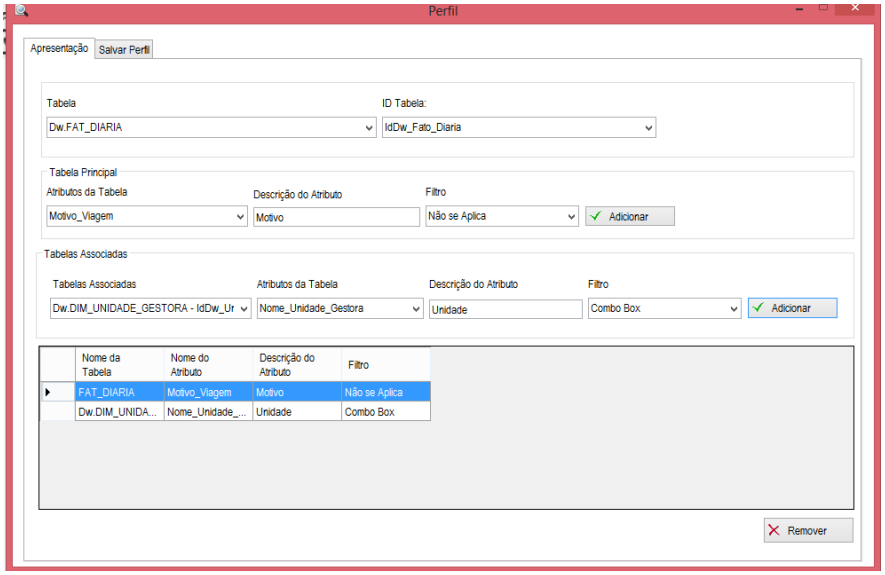

Figura 06. Tela Perfil (Ferramenta TextMining).

A criação de um perfil poderá ocorrer só uma vez e pode ser compartilhado por todos os usuários. Como o custo da operação é muito baixo, se houver a necessidade de alteração dele, basta excluí-lo e criar outro novamente. Esta característica torna a aplicação flexível e genérica através da geração de perfis de consulta diferenciados para qualquer tabela e campos contidos na base de dados.

Dados estes entendimentos sobre perfis, outra funcionalidade importantíssima é o gerenciamento de dicionários, que são os modelos de conhecimentos que servem de base para tornar possível a descoberta de evidências de fraudes semelhantes em toda base de dados ou em unidades e cidades específicas. Um dicionário é criado por meio da seleção de amostras que são dados selecionados pelo auditor como “Evidência” (possível evidência de irregularidade) e “Em Conformidade” (descrição que está de acordo com a lei), bem como o auditor pode informar amostras avulsas, as quais são especificadas manualmente e classificadas como “Evidência” ou “Em Conformidade”, como é mostrado na Figura 07.

A seleção de amostras para criação do dicionário deve ser balanceada, para cada evidência informada, deverá existir um ou mais registros que são exemplos de conformidade. Na ferramenta, está disponível para o usuário as funcionalidades de cadastro, consulta, edição, exclusão e desbloqueio de dicionários. É importante ressaltar que o dicionário criado poderá ser utilizado por todos os auditores, permitindo maior eficiência ao processo de auditoria.

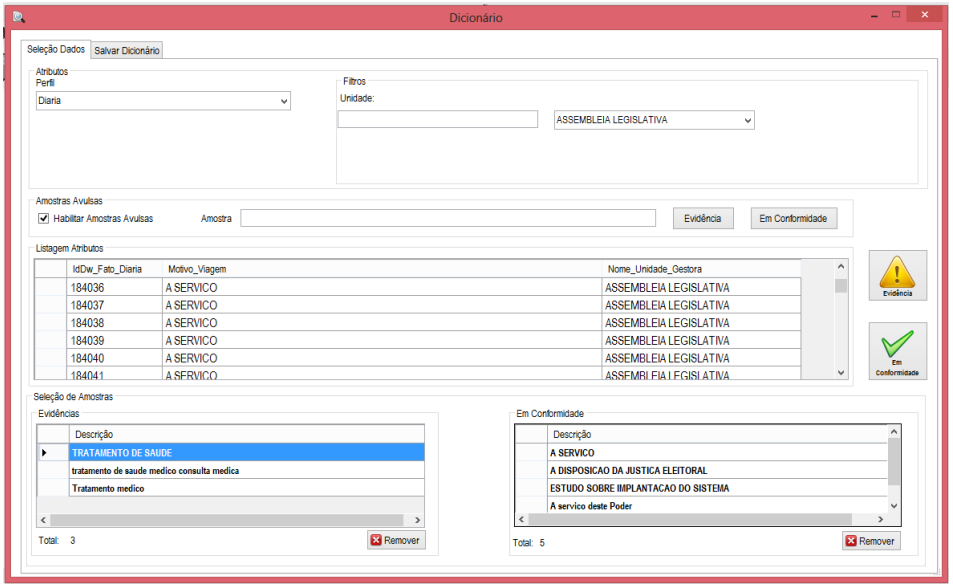

Figura 07. Tela Dicionário (Ferramenta TextMining).

A partir do perfil selecionado, dos filtros anexados a esse e do dicionário escolhido, o auditor poderá escolher os dados a serem classificados pela ferramenta, ou seja, local em que será buscado novas evidências semelhantes às do dicionário criado.

Durante a realização deste trabalho, a aplicação dispôs de dois algoritmos de mineração de texto, Naïve Bayes e Similaridade, para classificar os registros, como é mostrado na Figura 08. Ambos foram escolhidos mediante pesquisa bibliográfica sobre o uso em campos descritivos (texto).

Naïve Bayes é um algoritmo de análise estatística[3] e foi implementado de forma parametrizada, especificando um limiar mínimo para auxiliar na classificação dos registros. Para realizar a classificação de um registro, o algoritmo calcula a probabilidade desse registro ser ou não uma evidência de irregularidade. Este algoritmo dispõe de três formas para realizar o cálculo da probabilidade: “Híbrido”, “Frequência Inversa” e “Frequência”. Na primeira abordagem, é considerada a frequência do termo na amostra e o *tf* desse na sentença. Já na segunda, é levado em conta o *tfidf* do termo na amostra e o *tf* na sentença. Por fim, na terceira, são consideradas as frequências do termo na amostra e na sentença.

Já o algoritmo de Similaridade, também de análise estatística, calcula a similaridade entre uma sentença e um conjunto de amostras, por meio dos termos que ambos possuem em comum para determinar se a sentença é ou não uma evidência.

[3] Análise Estatística é uma das abordagens para análise de dados textuais, em que se leva em consideração a frequência dos termos no texto. Diferente da análise semântica que se baseia na sequência dos termos para determinar a função do termo no texto (MORAIS; AMBRÓSIO, 2007).

Na ferramenta, está disponível para o usuário as funcionalidades de cadastro, consulta, exclusão e visualização de classificações.

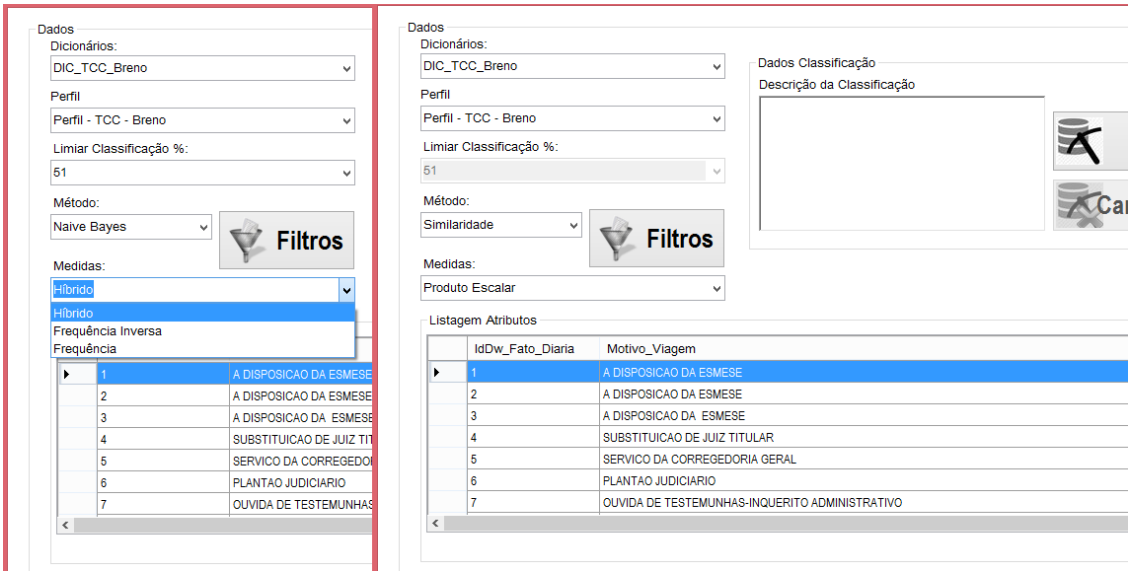

Figura 08. Telas Classificação sobrepostas mostrando os algoritmos disponíveis (Ferramenta TextMining).

3.2 Alterações e Melhorias na ferramenta TextMining

Após a análise do código da aplicação, foram efetuadas alterações no código, objetivando melhoria no uso da ferramenta, inclusão de novas funcionalidades, prevenção e correção de problemas. Abaixo seguem as principais alterações realizadas:

- **Modelo de Dados:**
  - Inclusão dos atributos “Metodo_Classificacao” e “Tempo_Classificacao” na tabela “DIM_CLASSIFICACAO”.
- **Módulo Dicionário:**
  - A tela de criação de dicionários foi alterada para permitir a inclusão de amostras avulsas;
  - A tela de consulta de dicionários foi alterada para que os botões “Editar” e “Excluir” ficassem desabilitados quando não existissem dicionários cadastrados;
  - A tela “Dicionários Bloqueados” foi alterada para que o botão “Liberar” ficasse desabilitado quando não existissem dicionários bloqueados.

- **Módulo Classificação:**
    - Criação da classe Similaridade, algoritmo de classificação;
    - Criação da classe abstrata Classificador, super-classe das classes Similaridade e NaiveBayes. A classe Classificador possui um método estático ClassificadorFactory que retorna um objeto do tipo Classificador, o qual pode ser uma instância das classes NaiveBayes ou Similaridade;
    - Na tela "Classificação", o algoritmo "Similaridade" foi incluído nas opções de métodos de classificação;
    - Na tela "Classificação", foram adicionados os percentuais 51 e 55 ao componente "Limiar Classificação %";
    - A tela "Classificação" foi alterada para os componentes "Dicionários", "Perfil", "Limiar Classificação %", "Método", "Medidas", "Filtros", "Classificar" e "Descrição da Classificação" serem desabilitados quando um processo classificatório fosse iniciado;
    - A tela de consulta de classificações foi alterada para que os botões "Excluir" e "Detalhes" ficassem desabilitados quando não existissem classificações cadastradas;
    - As alterações da tela "Dados da Classificação" foram:
        - Criação do componente "Tempo de Classificação" para visualizar o tempo da classificação realizada;
        - Atribuição do valor "---" para o componente "Limiar de Classificação %" quando o algoritmo de mineração utilizado não for o Naïve Bayes, porque o algoritmo de Naïve Bayes é o único que utiliza limiar;
        - Parametrização da tabela "Evidências", em que o label e os valores da coluna do Limiar/Score serão formatados de acordo com o algoritmo utilizado na classificação. Por exemplo, caso o algoritmo seja Naïve Bayes, o label da coluna será "Limiar" e os valores da coluna estarão formatados em porcentagem, mas se for escolhido

Similaridade, o label da coluna será “Score” e os valores da coluna estarão formatados em números com casas decimais.

3.3 Algoritmo Implementado

Neste trabalho, foi implementada uma adaptação do algoritmo de similaridade de documentos. O método convencional de similaridade, conforme foi visto na seção 2.3, realiza um cálculo de similaridade entre todos os documentos do dicionário e o documento a ser classificado, apenas levando em conta os termos que ocorrem em ambos.

Para classificar o novo documento, o algoritmo convencional apenas considera as ocorrências do maior *score*, ignorando todos os outros. Nessa abordagem, existe a possibilidade de não classificar um novo documento quando a quantidade de documentos com maior *score*, para diferentes classes, é a mesma, como é mostrado na Figura 09.

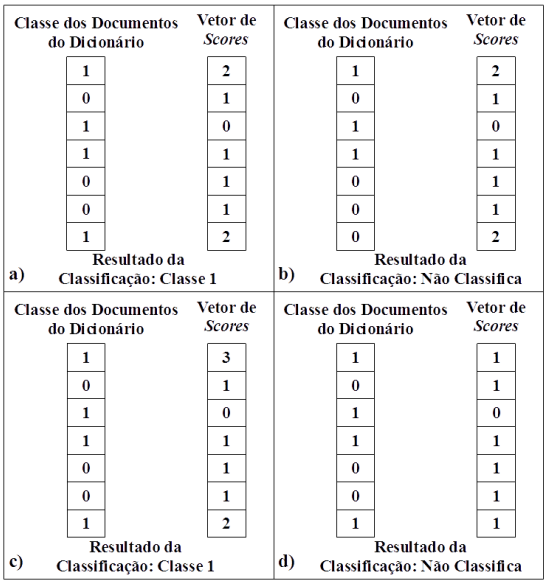

Figura 09. Exemplo de quatro situações de classificação do algoritmo convencional de similaridade.

Visando performance e melhoria do processo de classificação, foi desenvolvido uma adaptação do algoritmo de similaridade. As adaptações efetuadas foram:

- Será realizado uma poda, onde não serão analisados todos os documentos (amostras) do dicionário, somente aqueles que possuírem pelo menos um termo do documento (sentença) a ser classificado para determinar o grau de similaridade;
- Serão utilizados os outros *scores* quando não for possível classificar com o maior *score*.

O *score* para uma determinada amostra (documento do dicionário) pode ser obtido através do somatório da multiplicação do *tfidf* do termo da amostra com o *tf* da sentença (novo documento) para todos os termos comuns entre a amostra e a sentença, como pode ser visto na Fórmula 13, onde *k* é o número de termos que a amostra e a sentença possuem em comum.

$$score = \sum_{j=1}^{k} \left(tfidf_{amostra}(j) \times tf_{sentença}(j) \right)$$

Fórmula 13. Fórmula para o cálculo do *score* utilizado no algoritmo implementado.

Para facilitar o entendimento do algoritmo implementado, é apresentado o seu passo-a-passo (Quadro 01). No Apêndice A, é apresentado o seu pseudocódigo.

Entrada:
sc, sentença a ser classificada como "Evidência" ou "Em Conformidade"
dic, dicionário utilizado para classificar a sentença

Saída:
result, resultado da classificação da sentença, ou seja, objeto ResultadoGenerico contendo a classe e o score da sentença classificada

1. Calcular o *tf* (*term frequency*) para cada termo de *sc*.
2. Para cada amostra que contenha pelo menos um termo de *sc*, calcular o *score* para cada amostra, criar um objeto ResultadoGenerico para cada amostra analisada, contendo nesse objeto a classe da amostra e o *score* calculado e, por fim, armazenar o objeto no vetor de ResultadoGenerico.
3. Ordenar o vetor de ResultadoGenerico em ordem crescente pelo *score*.
4. **Se** o vetor estiver vazio **Então**
    Retorne um objeto ResultadoGenerico com classe igual a falso e *score* igual a zero.
Senão
5. **Se** o vetor possuir apenas um elemento **Então**
    Retorne o único objeto ResultadoGenerico dentro do vetor.
Senão
6. **Se** dentre os objetos ResultadoGenerico no vetor existe apenas uma ocorrência de um objeto com o maior *score* **Então**
    Retorne o objeto ResultadoGenerico com maior *score*.
Senão
7. Para todas as ocorrências dos objetos ResultadoGenerico com maior *score*, realizar a contagem de objetos que possuem classe igual a falso ("Em Conformidade") e também os que possuem classe igual a verdadeiro ("Evidência").
8. **Se** a quantidade de verdadeiro for maior que a de falso **Então**
    Retorne um objeto ResultadoGenerico com classe igual a verdadeiro e *score* igual ao maior *score*.
Senão
9. **Se** a quantidade de falso for maior que a de verdadeiro **Então**
    Retorne um objeto ResultadoGenerico com classe igual a falso e *score* igual ao maior *score*.
Senão
10. Enquanto não for possível classificar a sentença (quantidade de verdadeiros e falsos forem iguais) e nem todos os objetos do vetor foram analisados, realizar os passos 6 a 9, considerando que o novo maior *score* será o *score* do objeto ResultadoGenerico que antecede a primeira ocorrência do objeto que possui o atual maior *score*.
11. **Se** não foi possível classificar analisando todos os elementos do vetor **Então**
    Retorne um objeto ResultadoGenerico com classe igual a falso e *score* igual a zero.

Quadro 01. Passo-a-passo do algoritmo implementado.

4 ESTUDO DE CASO

Neste capítulo, serão apresentadas as atividades necessárias para a execução do estudo de caso realizado. Na seção 4.1, o objetivo do estudo realizado é apresentado. Em seguida, na 4.2, o planejamento do estudo é abordado, em que foram selecionados os participantes e objetos, bem como a definição do dicionário e métricas de desempenho e qualidade utilizadas. Por fim, na 4.3, é apresentado o processo de operação, o qual consiste na execução do estudo de caso.

4.1 Definição de Objetivo

A realização do estudo de caso tem por objetivo principal a validação dos resultados emitidos pela ferramenta TextMining para detecção de irregularidades nos pagamentos de diárias contidos nos históricos de contas públicas sob custódia do TCE-SE. Para atingir este objetivo, é necessária a efetivação dos seguintes objetivos específicos:

- Selecionar os participantes e objetos do estudo de caso;
- Definir o dicionário a ser utilizado;
- Executar o processo classificatório nas amostras dos participantes envolvidos para cada algoritmo de mineração de texto;
- Verificar e validar os resultados obtidos por meio das métricas de Tempo Médio de Execução, Acurácia, Precisão, Cobertura e Medida F;
- Realizar alterações na ferramenta, se necessários.

Após a definição dos objetivos, o planejamento a ser executado é apresentado abaixo.

4.2 Planejamento

Para garantir o alcance dos objetivos definidos, torna-se necessária a definição de uma estratégia de execução. Primeiramente, serão selecionados os participantes e objetos, em seguida, a definição do dicionário utilizado, a determinação das métricas para a avaliação de desempenho e qualidade dos algoritmos e, por fim, de acordo com a seção 4.3, a operação de execução.

4.2.1 Seleção de Participantes e Objetos

Para a seleção dos participantes, é necessário analisar dois critérios: os participantes devem ser unidades gestoras cadastradas no SISAP e que possuam uma quantidade considerável de registros na Tabela de Fatos[4] de Diárias. De acordo com o DW cedido, existem 481 unidades gestoras cadastradas, sendo assim, serão escolhidas, aleatoriamente, três unidades para a realização do estudo. Por questão de sigilo das informações do TCE-SE, os nomes das unidades gestoras não serão revelados.

As unidades escolhidas, com a quantidade de registros na Tabela de Fatos de Diárias especificados entre parênteses, foram: Unidade A (8872), Unidade B (625) e Unidade C (1855). É importante ressaltar que para as unidades A e C, também serão escolhidas dentro da quantidade de registros, aleatoriamente, amostras de 500 registros para o estudo. É fundamental frisar que a base de treinamento será constituída pela Unidade A, já a base de teste será formada pelas unidades B e C.

Após a escolha das unidades, é fundamental determinar o atributo na tabela de fato a ser minerado, ou seja, o campo descritivo. De acordo com a Figura 10, existem cinco campos descritivos: IdDw_Fato_Diaria, Matricula_Funcionario, Destino_Viagem, Motivo_Viagem e Numero_Empenho. Dentre estes, para detectar irregularidades no pagamento de diárias, o atributo mais significativo é Motivo_Viagem, porque o mesmo representa a justificativa da concessão de uma diária.

[4] Tabela de Fato: Também chamada de Tabela Dominante, é uma tabela que compõe o modelo multidimensional (Esquema Estrela) em que são armazenados grande quantidade de dados históricos, bem como os indicadores de desempenho (métricas) do negócio (COLAÇO JÚNIOR, 2004).

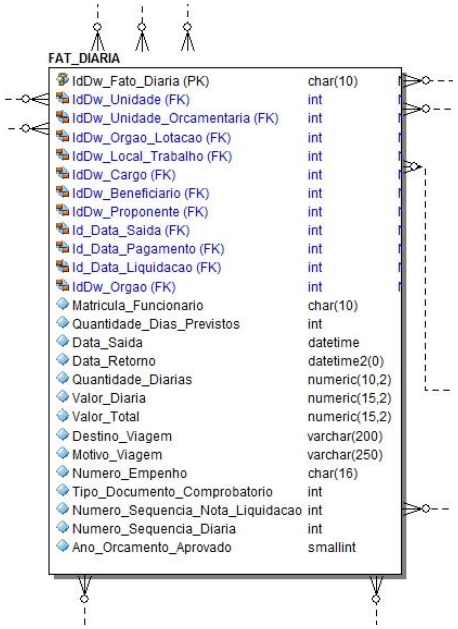

Figura 10. Tabela de Fato de Diárias (Modelo de Dados do DW do SISAP).

4.2.2 Dicionário Utilizado

De acordo com o Decreto Nº. 12.424, de 12/09/1991, do Governo de Sergipe, diária é uma espécie de auxílio financeiro ou ajuda de custo para um colaborador prestar algum serviço fora da localidade do órgão ao qual esteja vinculado. Em outras palavras, é um auxílio recebido pelo colaborador com o intuito de custear seus gastos para a realização de serviço fora do local de trabalho. A concessão de diárias é diversificada, pois abrange gastos referentes à capacitação, viagens para reuniões com superiores, entre outras.

Diante do exposto, é proibida a concessão de diárias para fins que não sejam relacionados à prestação de serviço. Existem inúmeras justificativas consideradas evidências de irregularidades para concessão de diárias como, por exemplo, realização de uma viagem particular. Assim, para a definição do dicionário a ser utilizado neste estudo, optou-se em restringir esse conjunto para evidências relacionadas ao uso de diárias para tratamento de saúde.

O modelo de conhecimento (dicionário a ser utilizado) possuirá, no total, 60 sentenças constituídas de amostras da própria base e de amostras avulsas para classificar registros. Para as amostras da própria base, foram escolhidas, aleatoriamente, 40 sentenças da Unidade A, sendo 20 classificadas como "Evidência" e as outras 20 como "Em Conformidade".

Para as amostras avulsas, foram definidas 20 sentenças, sendo 10 classificadas como "Evidência" e as outras 10 como "Em Conformidade". As sentenças avulsas classificadas como "Em Conformidade" são similares às da própria base, as quais foram formuladas por meio de uma análise das amostras dos dados das unidades gestoras envolvidas.

Já as sentenças avulsas classificadas como "Evidência", apenas duas são similares às da própria base, por meio da análise da amostra dos dados da Unidade A. Para formular as sentenças avulsas restantes, com o intuito de obter termos da Medicina sobre procedimentos, tratamentos e cirurgias médicas, profissionais da saúde, doenças e exames, foram realizadas algumas pesquisas sobre "Medicina" no site Wikcionário (http://pt.wiktionary.org/) e "Lista de Doenças" e "Lista de Doenças causadas por Seres Vivos" no site Wikipédia (http://pt.wikipedia.org), bem como no site boaSAÚDE (http://www.boasaude.com.br) foi obtida uma lista com 218 tipos de exames de rotina.

Com a ajuda de uma especialista na área da saúde, mestranda em Ciências da Saúde da UFS, foram selecionados, por categoria, apenas os termos mais comuns e relevantes, a exemplo de doenças mais comuns e exames de rotinas mais solicitados, como pode ser visto no Apêndice B. Vale destacar que foi verificada a existência desses termos em dois dicionários da área da saúde: Compacto Dicionário Ilustrado de Saúde (SILVA, 2007) e Dicionário de Termos Médicos e de Enfermagem (GUIMARÃES, 2002).

Nas tabelas 03 e 04, são apresentadas as amostras que constituem o modelo de conhecimento.

Tabela 03. Amostras da Própria Base (DW do SISAP).

AMOSTRAS DA PRÓPRIA BASE	
Sentença – "Em Conformidade"	**Sentença – "Evidência"**
A DISPOSICAO DA JUSTICA ELEITORAL	ACOMPANHAMENTO DE TRATAMENTO DE SAUDE DE SUA FILHA
A SERVIÇO DA ASSEMBLÉIA	ACOMPANHANTE DA SRA DEP. CELIA FRANCO P/TRATAMENTO MEDICO
A SERVIÇO DESTE PODER	ACOMPANHAR A DEPUTADA PARA TRATAMENTO DE SAUDE
A TRABALHO	ACOMPANHAR A FILHA EM TRATAMENTO MÉDICO
ACOMPANHAR O SR.PRIMEIRO SECRETARIO	ATENDER PROCEDIMENTOS MÉDICOS
ACOMPANHAR PROCESSOS E REUNIÕES PARLAMENTARES	CONSULTA MEDICA
ASSUNTO DE INTERESSE DESTE PODER	DESPESAS MÉDICA
AUTORIZADA PELO PRIMEIRO SECRETÁRIO	FAZER EXAMES PARA LIBERAÇÃO DE TRANSPLANTE DE RINS
ENCONTRO DO PARTIDO PROGRESSISTA	PARA A FUNCIONARIA A TRATAMENTO DE SAUDE
ESTUDO SOBRE IMPLANTACAO DO SISTEMA	PARA A SERVIDORA REALIZAR TRATAMENTO CLÍNICO
INTEGRAR COMITIVA DO GOVERNO DO ESTADO	REVISÃO MEDICA
PARA O SR. DEPUTADO PARTICIPAR DE REUNIÃO DO PARTIDO	SUBMETER-SE A CONSULTA MÉDICA
PARTCIPAR DE REUNIÃO DE CUNHO POLÍTICO PARTIDÁRIO	SUBMETER-SE A EXAMES MEDICOS
PARTICIPAR DA CONVENÇÃO NACIONAL DO DEM	SUBMETER-SE A TRATAMENTO MÉICO
PARTICIPAR DA POSSE DO PRESIDENTE DA PETROBRAS DISTRIBUIDORA	TRASTAMENTO MEDICO
PARTICIPAR DE ATO PUBLICO CONTRA REFORMA SINDICAL	TRATAMEMNTO DE SAUDE
PARTICIPAR DO I CONGRESSO INTERMUNICIPAL DE SAUDE	TRATAMENTI DE SAÚDE
REUNIÃO DE CUNHO POLITICO PARTIDARIO	TRATAMENTO DSE SAUDE
TRATAR DE ASSUNTO DE INTERESSE DESTE PODER	TRATAMNETO DE DE SAUDE
VISITAR A SUPERINTENDENCIA DA CAIXA ECONOMICA FEDERAL	TRATAMNETO DE SAUDE

Tabela 04. Amostras Avulsas.

AMOSTRAS "AVULSAS"	
Sentença – "Em Conformidade"	**Sentença – "Evidência"**
COMPLEMENTAÇÃO DE DIÁRIA PARA FUNCIONÁRIO REALIZAR TRABALHO.	25-HIDROXIVITAMINA D OU 25(OH)D; ACIDO ÚRICO NO SANGUE; ALBUMINA; ALTERAÇÕES DO FERRO E DE SUA CAPACIDADE DE FIXAÇÃO; AUDIOMETRIA VON BEKESY; BILIRRUBINA NA URINA; BILIRRUBINA NO SANGUE (DIRETA, INDIRETA E TOTAL). PROVAS DE FUNÇÃO HEPÁTICA (BILIRRUBINAS, ELETROFORESE DE PROTEÍNAS. FA, TGO, TGP E GAMA-PGT); CÁLCIO NO SANGUE; CARDIOLIPINA, AUTO-ANTICORPOS IGG; CITOGENÉTICA DIAGNÓSTICO PRÉ-NATAL; CLEARANCE DE URÉIA; CLEARANCE DE CREATININA; CLORO NO SANGUE (CL); COLESTEROL TOTAL; COLONOSCOPIA; CORTISOL PLASMÁTICO; CREATININA NO SANGUE; CREATINOFOSFOQUINASE OU CPK; CULTURA BACTERIOLÓGICA DO SANGUE (HEMOCULTURA); CURVA DE TOLERÂNCIA A GLICOSE; DENGUE, SOROLOGIA; DENSITOMETRIA ÓSSEA; DIAGNÓSTICO LABORATORIAL DA HEPATITE; DOPPLER SCAN COLORIDO ARTERIAL DE MEMBRO INFERIOR E SUPERIOR, COLORIDO DE VÍSCERAS ABDOMINAIS, DE CARÓTIDAS E VERTEBRAIS, VENOSO DE MEMBRO INFERIOR - UNILATERAL; ECOCARDIOGRAFIA, ECODOPPLERCARDIOGRAMA TRANSTORÁCICO; ELETROCARDIOGRAMA (ECG); ENDOSCOPIA; EXAME PARASITOLÓGICO DE FEZES; FERRITINA NO SANGUE; FERRO SÉRICO; FIBRINOGÊNIO PLASMÁTICO; FIBROSE CÍSTICA, ESTUDO GENÉTICO; FOSFATASE ÁCIDA, ALCALINA, ALCALINA NEUTROFÍLICA OU LEUCOCITÁRIA; GAMA-GLUTAMIL TRANSFERASE (GGT); GLICEMIA PÓS PRANDIAL; HIV AIDS (SÍNDROME DE IMUNO DEFICIÊNCIA ADQUIRIDA) (EXAME DE WESTERN BLOT E TESTE DE ELISA); HPV CAPTURA HÍBRIDA PROCEDIMENTO DIAGNÓSTICO POR CAPTURA HÍBRIDA; HEMOGRAMA COM CONTAGEM DE PLAQUETAS OU FRAÇÕES (ERITROGRAMA, ERITRÓCITOS, LEUCÓCITOS, LEUCOGRAMA, PLAQUETAS); HEMOSSEDIMENTAÇÃO; HORMÔNIO DE CRESCIMENTO NO SANGUE. HORMÔNIO SOMATOTRÓFICO (STH); HORMÔNIO LUTEINIZANTE NO PLASMA; HORMÔNIO PARATIREOIDEANO NO SANGUE; IMUNOGLOBULINAS E TOTAL, G, A E M NO SANGUE; INSULINA NO SANGUE; MAGNÉSIO NO SANGUE (MG+); MAMOGRAFIA;

	MICROALBUMINÚRIA; PAPANICOLAU (CITOLOGIA VAGINAL); PEPTÍDEO C; POTÁSSIO NO SANGUE (K+); PROTEÍNA C REATIVA; RAIO X DA PERNA, DO ANTEBRAÇO, DO BRAÇO, DOS SEIOS DA FACE; RESSONÂNCIA MAGNÉTICA (RM) DE CRÂNIO (ENCÉFALO), DA COLUNA; SANGUE OCULTO NAS FEZES, PESQUISA; TSH; TEMPO DE COAGULAÇÃO E DE RETRAÇÃO DO COÁGULO; TESTE ERGOMÉTRICO; TESTOSTERONA LIVRE; TIROXINA (T4); TOMOGRAFIA COMPUTADORIZADA (TC) DE ABDOMEM, DE COLUNA VERTEBRAL, DE CRÂNIO, DE TÓRAX, DOS SEIOS PARANASAIS; TRANSAMINASE OXALACÉTICA (TGO), PIRÚVICA (TGP); TRANSFERRINA; TRI IODO TIRONINA (T3); TRIGLICÉRIDES; ULTRASSONOGRAFIA, ULTRA-SONOGRAFIA (US), ULTRASSOM ABDOMINAL ABDOME INFERIOR MASCULINO OBSTÉTRICA (BEXIGA, PRÓSTATA E VESÍCULAS SEMINAIS) ABDOME INFERIOR FEMININO (BEXIGA, ÚTERO, OVÁRIO E ANEXOS) ABDOME TOTAL (INCLUI PELVE) ABDOME SUPERIOR (FÍGADO, VIAS BILIARES, VESÍCULA, PÂNCREAS, BAÇO), DA TIREÓIDE, DA MAMA; URINA (ANÁLISE DE ROTINA); UROCULTURA; URÉIA NO SANGUE (NITROGÊNIO UREICO)
CONDUZIR PACIENTES PARA HOSPITAL.	ABLATIVA; ABORTO; ACUPUNTURA; ALOPATIA; AMPUTAÇÃO; ANESTESIA; ANTI-SEPSIA; APENDICECTOMIA; ASSEPSIA; AUSCULTAÇÃO; AUTÓPSIA; BARIÁTRICA; BIÓPSIA; CABEÇA; CARDÍACA; CAUTERIZAÇÃO; CHECK-UP; CIRURGIA; COLUNA; COSTURA; DRENO; ELETROCIRURGIA; EXAME; HEMODIÁLISE; HERNIOTOMIA; HIDROTERAPIA; HISTERECTOMIA; HOMEOPATIA; IMPLANTE; LAPAROSCOPIA; LAVAGEM; LIPOASPIRAÇÃO; LOBOTOMIA; MASSAGEM; MASTECTOMIA; NEFRECTOMIA; NEUROCIRURGIA; OBTURAÇÃO; OCLUSÃO; OCUPACIONAL; OPERAÇÃO; OPERAÇÃO CESARIANA OU CESÁREA; ORTOPÉDICA; PESCOÇO; PLÁSTICA; PROFILÁTICA; PSICANÁLISE; PUNÇÃO; QUIMIOTERAPIA; QUIROPATIA; QUIROPRAXIA; RADIOCIRURGIA; RADIOSCOPIA; RADIOTERAPIA; RINOTOMIA; SONOTERAPIA; SOROTERAPIA; TERAPIA; TRANSFUSÃO; TRANSFUSÃO DE SANGUE; TRANSPLANTAÇÃO; TRANSPLANTE; TRAQUEOPLASTIA; TRATAMENTO; TRATAMENTO DE CHOQUE; TREPANAÇÃO; ULTRA-SONOCIRURGIA; VACINAÇÃO;

	VASECTOMIA; VIDEOCIRURGIA; ZONULOTOMIA
CONDUZIR VÍTIMAS DE ABUSO SEXUAL PARA TRATAMENTO CLÍNICO E PSICOLÓGICO.	ABSCESSO; ALERGIA; APENDICITE; ASFIXIA; BACTÉRIA; CRISE; CÁLCULO RENAL; DERRAME; DISFAGIA; DISFUNÇÃO; DISTENSÃO; DISTROFIA; DOENÇA; EDEMA; ENXAQUECA; ESCORIAÇÃO; ESPASMO; ESTIRAMENTO; FRATURA; FUNGO; HANSENÍASE; HEMORRAGIA; INFARTO; INFECÇÃO; INFLAMAÇÃO; INSOLAÇÃO; INSUFICIÊNCIA; INSÔNIA; LESÃO; MUDEZ; OBESIDADE; PEDRA NO RIM; PNEUMONIA; PROTOZOÁRIO; REAÇÃO; RECORRÊNCIA; REJEIÇÃO; REUMATISMO; SEQUELA; SINTOMA; STRESS; SUFOCAMENTO; SUFOCAÇÃO; SÍNCOPE; SÍNDROME; TORCICOLO; TRAUMA; TRAUMATISMO; VERME; VERMINOSE; VÍRUS
PAGAMENTO DE DIÁRIA AO MOTORISTA PARA CONDUZIR A COMITIVA DO MINISTRO DA SAÚDE.	ACNE; ANEMIA; ANSIEDADE; ARTRITE; ARTROSE; ASCARIDÍASE; ASMA; BERIBÉRI; CANCRO, TUMOR OU CÂNCER; CIRROSE HEPÁTICA; CÁRIE; DENGUE; DEPRESSÃO; DERMATITE SEBORRÉICA, SEBORRÉIA OU CASPA; DERMATOFITOSE, MICOSE; DIABETES INSIPIDUS MELLITUS; DIARREIA; DISENTERIA AMÉBICA OU AMEBIANA, AMEBÍASE; DISENTERIA BACTERIANA OU SHIGELOSE; DISLIPIDEMIA; DOENÇA DE CHAGAS, CHAGUISMO OU TRIPANOSSOMÍASE AMERICANA; ÉBOLA; EPILEPSIA; ESCABIOSE OU SARNA; ESCLEROSE MÚLTIPLA; ESOFAGITE; ESQUISTOSSOMOSE OU BILHARZÍASE; FARINGITE; FEBRE; FIBROSE CÍSTICA; GASTRITE; GIARDIOSE OU GIARDÍASE; GLAUCOMA; GOTA; HEPATITE; HERPES; HIPERCOLESTEROLEMIA; HIPERPARATIROIDISMO; HIPERTENSÃO ARTERIAL OU PULMONAR; HIPERTIROIDISMO; HIPOTIROIDISMO; LEISHMANIOSE, LEISHMANÍASE, CALAZAR OU ÚLCERA DE BAURU; LEUCEMIA MIELOIDE AGUDA; LEUCEMIA OU LINFOMA DE CÉLULAS T DO ADULTO; LÚPUS ERITEMATOSO SISTÊMICO; MAL DE ALZHEIMER; MAL DE PARKINSON; MALÁRIA OU PALUDISMO; MENINGITE; OSTEOPOROSE; PNEUMONIA; RUBÉOLA OU RUBELA; SARAMPO; SINUSITE; SÍFILIS; SÍNDROME DA IMUNODEFICIÊNCIA ADQUIRIDA, AIDS OU SIDA HIV; TRANSTORNOS ALIMENTARES; TUBERCULOSE; ÚLCERA; VARICELA OU CATAPORA
PAGAMENTO DE DIÁRIA PARA SERVIDOR OU FUNCIONÁRIO REALIZAR SERVIÇOS FORA DESTA UNIDADE.	AMBULATÓRIO; ASSISTÊNCIA MÉDICA; ATENDIMENTO MÉDICO; CARDIOGRAMA; CLÍNICA; CONSULTA MÉDICA; CONSULTÓRIO; DIAGNÓSTICO MÉDICO; DESPESA MÉDICA;

	ELETROENCEFALOGRAMA; EMERGÊNCIA; EXAMES MÉDICOS; HISTÓRICO DE SAÚDE; HOSPITAL; LAUDO; PERÍCIA MÉDICA; POLICLÍNICA; PROCEDIMENTO DE SAÚDE; PRONTO-SOCORRO; QUADRO CLÍNICO; RADIOGRAFIA; REVISÃO MÉDICA; VACINA
PARTICIPAR DE OFICINA, TREINAMENTO, CURSO, CAPACITAÇÃO, CONGRESSO, SEMINÁRIO, SIMPÓSIO, FÓRUM, CONVENÇÃO, ENCONTRO, FEIRA NA ÁREA DA SAÚDE.	ANDROLOGIA; ANESTESIOLOGIA; ANGIOLOGIA; AUXOLOGIA; BIOMEDICINA; CANCEROLOGIA; CARDIOLOGIA; DERMATOLOGIA; ENDOCRINOLOGIA; EPIDEMIOLOGIA; ESTOMATOLOGIA; FISIOTERAPIA; FONIATRIA; FONOAUDIOLOGIA; GASTRENTEROLOGIA; GERIATRIA; GERONTOLOGIA; GINECOLOGIA; IMUNOLOGIA; MASTOLOGIA; NEFROLOGIA; NEONATOLOGIA; NEUROLOGIA; NEURORRADIOLOGIA; NUTRIÇÃO; OBSTETRÍCIA; ODONTOLOGIA; OFTALMOLOGIA; ONCOLOGIA; OPTOMETRIA; ORTODONTIA; ORTOPEDIA; OTORRINOLARINGOLOGIA; PATOLOGIA; PEDIATRIA; PNEUMOLOGIA; PODOLOGIA; PROCTOLOGIA; PSICOLOGIA; PSICOTERAPIA; PSIQUIATRIA; RADIOLOGIA; REUMATOLOGIA; SEROLOGIA; SINTOMATOLOGIA; SOMATOLOGIA; TERAPÊUTICA; TRAUMATOLOGIA; UROLOGIA
PARTICIPAR DE UMA REUNIÃO COM SECRETÁRIO DA SAÚDE.	ANDROLOGISTA; ANESTESIOLOGISTA; ANESTESISTA; CARDIOLOGISTA; CARDIÓLOGO; CIRURGIÃO; CIRURGIÃO-DENTISTA; DENTISTA; DERMATOLOGISTA; DOUTOR; ENDOCRINOLOGISTA; ENDÓCRINO; ENFERMEIRA; EPIDEMIOLOGISTA; FISIOTERAPEUTA; FONOAUDIÓLOGO; GASTRENTEROLOGISTA; GERIATRA; GERONTOLOGISTA; GERONTÓLOGO; GINECOLOGISTA; HOMEOPATA; IMUNOLOGISTA; LEGISTA; MÉDICO; MÉDICO-LEGISTA; NEFROLOGISTA; NEFRÓLOGO; NEONATOLOGISTA; NEUROCIRURGIÃO; NEUROLOGISTA; NUTRICIONISTA; OBSTETRA; OFTALMOLOGISTA; OFTALMÓLOGO; ONCOLOGISTA; OPTOMETRISTA; ORTOPEDISTA; OSTEOPATA; OTORRINOLARINGOLOGISTA; PATOLOGISTA; PEDIATRA; PODÓLOGO; PROCTOLOGISTA; PSICANALISTA; PSICOTERAPEUTA; PSICÓLOGO; PSIQUIATRA; RADIOLOGISTA; REUMATOLOGISTA; SANITARISTA; SEROLOGISTA; TERAPEUTA; TERAPISTA; TRAUMATOLOGISTA; UROLOGISTA; URÓLOGO
PARTICIPAR DE UMA REUNIÃO, AUDIÊNCIA, CONFERÊNCIA, ATO PÚBLICO, ASSEMBLÉIA COM MINISTRO DA SAÚDE, GOVERNADOR, VICE-GOVERNADOR, PREFEITO, VICE-PREFEITO E SECRETÁRIOS.	CIRÚRGICO; CITOPATOLÓGICO; CLÍNICO; DERMATOLÓGICO; ECOCARDIOGRÁFICO; ELETRENCEFALOGRÁFICO; EPIDEMIOLÓGICO; FISIOLÓGICO; FISIOTERÁPICO; FONOAUDIOLÓGICO; FÍSICO;

	GERIÁTRICO; GERONTOLÓGICO; GINECOLÓGICO; HIPOCRÁTICO; HOMEOPÁTICO; HOSPITALAR; IDIOSSINCRÁTICO; IMUNITÁRIO; IMUNOLÓGICO; LABORATORIAL; MEDICINAL; MÉDICO-HOSPITALAR; NEFROLÓGICO; NEUROLÓGICO; OFTALMOLÓGICO; ONCOLÓGICO; OPERATÓRIO; OPTOMÉTRICO; ORTOPÉDICO; PARALÍTICO; PARAMÉDICO; PATOLÓGICO; POLICLÍNICA; PROCTOLÓGICO; PROFILÁTICO; PSIQUIÁTRICO; QUADRIPLÉGICO; QUIMIOTERÁPICO; QUIROPRÁTICO; RADIOGRÁFICO; RADIOLÓGICO; RADIOSCÓPICO; REUMATOLÓGICO; SEROLÓGICO; SINTOMATOLÓGICO; SINTOMÁTICO; SOMATOLÓGICO; TERAPÊUTICO; TRAUMATOLÓGICO; TRAUMÁTICO; UROLÓGICO
VIAGEM PARA REALIZAÇÃO DE SERVIÇO DESTA UNIDADE.	REALIZAÇÃO DE EXAMES MÉDICOS DA ESPOSA E FILHOS.
VIAGEM PARA TRATAR DE ASSUNTOS DA SAÚDE PÚBLICA E OBTER RECURSOS FINANCEIROS.	REALIZAR PROCEDIMENTO CIRÚRGICO.

4.2.3 Medidas de desempenho e qualidade para avaliação dos algoritmos

Para analisar o desempenho e qualidade dos algoritmos de mineração de texto em questão, Naïve Bayes e Similaridade, será utilizado o recurso Matriz de Confusão, bem como as métricas de Acurácia, Cobertura, Precisão e Medida F e Tempo de Execução.

De acordo com o contexto deste trabalho, devemos considerar quatro situações:

- **NSCCE:** Número de sentenças classificadas corretamente como "Evidência" (*True Positive*).
- **NSCCC:** Número de sentenças classificadas corretamente como "Em Conformidade" (*True Negative*).
- **NSCEE:** Número de sentenças classificadas erroneamente como "Evidência" (*False Positive*).
- **NSCEC:** Número de sentenças classificadas erroneamente como "Em Conformidade" (*False Negative*).

A matriz de confusão que contempla as situações acima podem ser vista na Tabela 05.

Tabela 05. Matriz de Confusão utilizada.

Classificação Correta	**Classificado como**	
	Evidência	**Em Conformidade**
Evidência	*NSCCE*	*NSCEC*
Em Conformidade	*NSCEE*	*NSCCC*

Com a matriz de confusão definida, podemos definir as métricas a serem utilizadas.

4.2.3.1 Acurácia

Acurácia é o percentual de sentenças classificadas corretamente pelo classificador. Nesse contexto, pode ser determinada pela Fórmula 14.

$$acurácia = \frac{NSCCE + NSCCC}{NSCCE + NSCCC + NSCEE + NSCEC}$$

Fórmula 14. Fórmula da Acurácia.

4.2.3.2 Cobertura

Cobertura é o percentual de evidências que foram classificadas corretamente como "Evidência". Nesse contexto, pode ser determinada pela Fórmula 15.

$$cobertura = \frac{NSCCE}{NSCCE + NSCEC}$$

Fórmula 15. Fórmula da Cobertura.

4.2.3.3 Precisão

Precisão é o percentual de sentenças classificadas como “Evidência” que são realmente evidências. Nesse contexto, pode ser determinada pela Fórmula 16.

$$precisão = \frac{NSCCE}{NSCCE + NSCEE}$$

Fórmula 16. Fórmula da Precisão.

4.2.3.4 Medida F

Medida F, também conhecida como Média Harmônica da Precisão e Cobertura, é a medida que combina a precisão e cobertura. Nesse contexto, pode ser determinada pela Fórmula 17.

$$MedidaF = \frac{2 \times precisão \times cobertura}{precisão + cobertura}$$

Fórmula 17. Fórmula da Medida F.

4.2.3.5 Tempo de Execução

Tempo de Execução é o tempo de duração de uma classificação, compreendida pela diferença entre o tempo de término e o tempo de início da classificação. Nesse contexto, pode ser determinada pela Fórmula 18, onde ΔT é o tempo de execução, T_f o tempo de término da classificação e T_i o tempo de início da classificação.

$$\Delta T = T_f - T_i$$

Fórmula 18. Fórmula do Tempo de Execução.

4.3 Operação

Definido o planejamento, é de suma importância estabelecer a operação para a realização do estudo de caso.

4.3.1 Execução

Esta etapa consistirá na realização do processo classificatório nas amostras dos participantes envolvidos para cada algoritmo de mineração de texto, utilizando o modelo de conhecimento definido na seção 4.2.2. Foram efetuadas três classificações nas amostras dos participantes envolvidos para cada algoritmo. É necessário frisar que para cada execução do Naïve Bayes foi utilizado cada método desse algoritmo (“Híbrido”, “Frequência Inversa” e “Frequência”), bem como o limiar de 51 % foi utilizado em todas as execuções do Naïve Bayes. Após o término do processo classificatório, as matrizes de confusão foram geradas a partir dos resultados obtidos das classificações efetuadas, assim como foram coletadas as métricas para cada algoritmo.

Após a realização do estudo de caso, no Capítulo 5 são apresentados os resultados obtidos, assim como a análise comparativa sobre todas as abordagens.

5 RESULTADOS

Nesta seção, serão apresentados os resultados obtidos a partir da coleta das métricas, bem como a análise comparativa dos algoritmos de mineração de texto em questão. A análise comparativa foi feita de duas maneiras: por unidade gestora e por métrica utilizada.

Após a realização do estudo de caso, discutido no Capítulo 4, foram coletados os valores das matrizes de confusão de cada execução para cada algoritmo e unidade escolhida. Com os valores das matrizes de confusão, foi possível coletar as métricas para avaliar todas as abordagens. Vale ressaltar que para um mesmo algoritmo e unidade gestora, a matriz de confusão foi a mesma para as três execuções (processos classificatórios). Nas tabelas 06 e 07, é apresentado um resumo dos valores das matrizes de confusão por algoritmo e unidade.

Tabela 06. Valores da Matriz de Confusão por Algoritmo e Unidade Gestora – Diagonal Principal.

Unidades	**Valores da Matriz de Confusão – Diagonal Principal**							
	NSCCE (TP)				***NSCCC (TN)***			
	N.B. F.I.	**N.B. H.**	**N.B. F.**	**SIM.**	**N.B. F.I.**	**N.B. H.**	**N.B. F.**	**SIM.**
Unidade A	136	137	137	137	363	363	363	362
Unidade B	0	1	1	0	613	603	603	588
Unidade C	3	3	3	2	457	449	449	430

* N.B.F.I.: Naïve Bayes – Frequência Inversa; N.B.H.: Naïve Bayes – Híbrido; N.B.F.: Naïve Bayes – Frequência; SIM.: Similaridade.

Tabela 07. Valores da Matriz de Confusão por Algoritmo e Unidade Gestora – Diagonal Secundária.

Unidades	**Valores da Matriz de Confusão – Diagonal Secundária**							
	NSCEE (FP)				***NSCEC (FN)***			
	N.B. F.I.	**N.B. H.**	**N.B. F.**	**SIM.**	**N.B. F.I.**	**N.B. H.**	**N.B. F.**	**SIM.**
Unidade A	0	0	0	1	1	0	0	0
Unidade B	11	21	21	36	1	0	0	1
Unidade C	40	48	48	67	0	0	0	1

* N.B.F.I.: Naïve Bayes – Frequência Inversa; N.B.H.: Naïve Bayes – Híbrido; N.B.F.: Naïve Bayes – Frequência; SIM.: Similaridade.

Inicialmente, foram analisados os resultados das três classificações realizadas para cada algoritmo na Unidade A. Conforme é visto nas tabelas 08, 09 e 10, os algoritmos Naïve Bayes – Híbrido (N.B.H.) e Naïve Bayes – Frequência (N.B.F.) são as melhores abordagens para essa unidade, pois possuem as melhores porcentagens de acurácia (100%), precisão (100%), cobertura (100%) e medida F (100%). Similaridade (SIM.) obteve um melhor desempenho do que os demais na métrica Tempo de Execução. Comparando Similaridade e Naïve Bayes – Frequência Inversa (N.B.F.I.), Similaridade supera esse nas métricas de Cobertura, Medida F e Tempo de Execução, mas ambos possuem a mesma porcentagem de acurácia. É importante verificar que o algoritmo Similaridade possui precisão inferior em relação às demais abordagens.

Tabela 08. Comparativo das métricas Acurácia e Precisão para cada algoritmo na Unidade A.

Execuções	Métricas de Desempenho e Qualidade							
	Acurácia				Precisão			
	N.B. F.I.	N.B. H.	N.B. F.	SIM.	N.B. F.I.	N.B. H.	N.B. F.	SIM.
1ª Exec.	99,80 %	100 %	100 %	99,80 %	100 %	100 %	100 %	99,28 %
2ª Exec.	99,80 %	100 %	100 %	99,80 %	100 %	100 %	100 %	99,28 %
3ª Exec.	99,80 %	100 %	100 %	99,80 %	100 %	100 %	100 %	99,28 %
Média	**99,80 %**	**100 %**	**100 %**	**99,80 %**	**100 %**	**100 %**	**100 %**	**99,28 %**

Tabela 09. Comparativo das métricas Cobertura e Medida F para cada algoritmo na Unidade A.

Execuções	Métricas de Desempenho e Qualidade							
	Cobertura				Medida F			
	N.B. F.I.	N.B. H.	N.B. F.	SIM.	N.B. F.I.	N.B. H.	N.B. F.	SIM.
1ª Exec.	99,27 %	100 %	100 %	100 %	99,63 %	100 %	100 %	99,64 %
2ª Exec.	99,27 %	100 %	100 %	100 %	99,63 %	100 %	100 %	99,64 %
3ª Exec.	99,27 %	100 %	100 %	100 %	99,63 %	100 %	100 %	99,64 %
Média	**99,27 %**	**100 %**	**100 %**	**100 %**	**99,63 %**	**100 %**	**100 %**	**99,64 %**

Tabela 10. Comparativo da métrica Tempo de Execução para cada algoritmo na Unidade A.

Execuções	Métricas de Desempenho e Qualidade			
	Tempo de Execução			
	N.B. F.I.	N.B. H.	N.B. F.	SIM.
1ª Exec.	135,88 s	133,49 s	146,79 s	84,78 s
2ª Exec.	135,24 s	139,28 s	153,68 s	83,39 s
3ª Exec.	134,64 s	134,81 s	156,64 s	82,22 s
Média	**135,25 s**	**135,86 s**	**152,37 s**	**83,46 s**

A próxima unidade a ser analisada foi Unidade B. Também foram analisados os resultados das três classificações realizadas para cada algoritmo na referida unidade. De acordo com as tabelas 11, 12 e 13, o algoritmo Similaridade obteve um melhor desempenho do que os demais apenas na métrica de Tempo de Execução. Já o Naïve Bayes – Frequência Inversa obteve um melhor desempenho na métrica Acurácia (98,08 %). Já Similaridade obteve a menor porcentagem de acurácia. É importante verificar que todas as abordagens tiveram um péssimo desempenho na métrica Precisão (valor abaixo de 50 %), mas Naïve Bayes – Híbrido e Naïve Bayes – Frequência tiveram desempenho melhor do que os demais.

Apesar dos péssimos resultados, Naïve Bayes – Frequência Inversa foi a melhor abordagem, pois o mesmo classificou, erroneamente, um número muito inferior de evidências do que os outros algoritmos, como é mostrado nas tabelas 06 e 07. Em outras palavras, comparando-se o resultado da soma entre *NSCCE* e *NSCCC* (soma da diagonal principal da matriz de confusão) de cada algoritmo, o resultado de Naïve Bayes – Frequência Inversa foi superior aos resultados das demais abordagens, portanto, Naïve Bayes – Frequência Inversa foi o algoritmo que apresentou melhor desempenho e qualidade na classificação das sentenças dessa unidade.

Tabela 11. Comparativo das métricas Acurácia e Precisão para cada algoritmo na Unidade B.

Execuções	Métricas de Desempenho e Qualidade							
	Acurácia				Precisão			
	N.B. F.I.	N.B. H.	N.B. F.	SIM.	N.B. F.I.	N.B. H.	N.B. F.	SIM.
1ª Exec.	98,08 %	96,64 %	96,64 %	94,08 %	0,00 %	4,55 %	4,55 %	0,00 %
2ª Exec.	98,08 %	96,64 %	96,64 %	94,08 %	0,00 %	4,55 %	4,55 %	0,00 %
3ª Exec.	98,08 %	96,64 %	96,64 %	94,08 %	0,00 %	4,55 %	4,55 %	0,00 %
Média	**98,08 %**	**96,64 %**	**96,64 %**	**94,08 %**	**0,00 %**	**4,55 %**	**4,55 %**	**0,00 %**

Tabela 12. Comparativo das métricas Cobertura e Medida F para cada algoritmo na Unidade B.

Execuções	Métricas de Desempenho e Qualidade							
	Cobertura				Medida F			
	N.B. F.I.	N.B. H.	N.B. F.	SIM.	N.B. F.I.	N.B. H.	N.B. F.	SIM.
1ª Exec.	0,00 %	100 %	100 %	0,00 %	∄	8,70 %	8,70 %	∄
2ª Exec.	0,00 %	100 %	100 %	0,00 %	∄	8,70 %	8,70 %	∄
3ª Exec.	0,00 %	100 %	100 %	0,00 %	∄	8,70 %	8,70 %	∄
Média	**0,00 %**	**100 %**	**100 %**	**0,00 %**	∄	**8,70 %**	**8,70 %**	∄

Tabela 13. Comparativo da métrica Tempo de Execução para cada algoritmo na Unidade B.

Execuções	Métricas de Desempenho e Qualidade			
	Tempo de Execução			
	N.B. F.I.	N.B. H.	N.B. F.	SIM.
1ª Exec.	256,01 s	268,49 s	276,44 s	159,55 s
2ª Exec.	259,67 s	258,26 s	266,54 s	158,14 s
3ª Exec.	254,64 s	271,76 s	291,13 s	159,16 s
Média	**256,77 s**	**266,17 s**	**278,04 s**	**158,95 s**

Encerrando a primeira análise, a próxima unidade a ser analisada foi Unidade C. Também foram analisados os resultados das três classificações realizadas para cada algoritmo nessa unidade. Conforme é visto nas tabelas 14, 15 e 16, Similaridade obteve um melhor desempenho do que os demais apenas na métrica de Tempo de Execução, bem como obteve os menores percentuais nas outras métricas. Já o Naïve Bayes – Frequência Inversa obteve as melhores porcentagens de acurácia, precisão, cobertura e medida F, seguido das abordagens Naïve Bayes – Híbrido e Naïve Bayes – Frequência. Sendo assim, o algoritmo Naïve Bayes – Frequência Inversa foi o melhor método de classificação para as sentenças dessa unidade. Contudo, é importante observar que, apesar do ótimo desempenho, Naïve Bayes – Frequência Inversa classificou, erroneamente, uma quantidade considerável de sentenças como "Evidência", como é mostrado na Tabela 07.

Tabela 14. Comparativo das métricas Acurácia e Precisão para cada algoritmo na Unidade C.

Execuções	Métricas de Desempenho e Qualidade							
	Acurácia				Precisão			
	N.B. F.I.	N.B. H.	N.B. F.	SIM.	N.B. F.I.	N.B. H.	N.B. F.	SIM.
1ª Exec.	92,00 %	90,40 %	90,40 %	86,40 %	6,98 %	5,88 %	5,88 %	2,90 %
2ª Exec.	92,00 %	90,40 %	90,40 %	86,40 %	6,98 %	5,88 %	5,88 %	2,90 %
3ª Exec.	92,00 %	90,40 %	90,40 %	86,40 %	6,98 %	5,88 %	5,88 %	2,90 %
Média	**92,00 %**	**90,40 %**	**90,40 %**	**86,40 %**	**6,98 %**	**5,88 %**	**5,88 %**	**2,90 %**

Tabela 15. Comparativo das métricas Cobertura e Medida F para cada algoritmo na Unidade C.

Execuções	Métricas de Desempenho e Qualidade							
	Cobertura				Medida F			
	N.B. F.I.	N.B. H.	N.B. F.	SIM.	N.B. F.I.	N.B. H.	N.B. F.	SIM.
1ª Exec.	100 %	100 %	100 %	66,67 %	13,04 %	11,11 %	11,11 %	5,56 %
2ª Exec.	100 %	100 %	100 %	66,67 %	13,04 %	11,11 %	11,11 %	5,56 %
3ª Exec.	100 %	100 %	100 %	66,67 %	13,04 %	11,11 %	11,11 %	5,56 %
Média	**100 %**	**100 %**	**100 %**	**66,67 %**	**13,04 %**	**11,11 %**	**11,11 %**	**5,56 %**

Tabela 16. Comparativo da métrica Tempo de Execução para cada algoritmo na Unidade C.

Execuções	Métricas de Desempenho e Qualidade			
	Tempo de Execução			
	N.B. F.I.	N.B. H.	N.B. F.	SIM.
1ª Exec.	184,19 s	201,97 s	245,13 s	117,37 s
2ª Exec.	212,91 s	185,20 s	311,77 s	114,61 s
3ª Exec.	187,70 s	194,69 s	256,87 s	119,11 s
Média	**194,93 s**	**193,95 s**	**271,26 s**	**117,03 s**

Encerrando a análise comparativa dos algoritmos, foi feita uma análise por métrica utilizada.

Verificando o Gráfico 01, correspondente à métrica Acurácia, é notável os ótimos desempenhos dos Naïve Bayes – Híbrido e Naïve Bayes – Frequência na Unidade A e o empate de ambos em todas unidades gestoras. Também é perceptível o empate entre Naïve Bayes – Frequência Inversa e Similaridade na Unidade A. Em média, Naïve Bayes – Frequência Inversa possui a melhor porcentagem de Acurácia, consequentemente, é a melhor abordagem em termos de Acurácia.

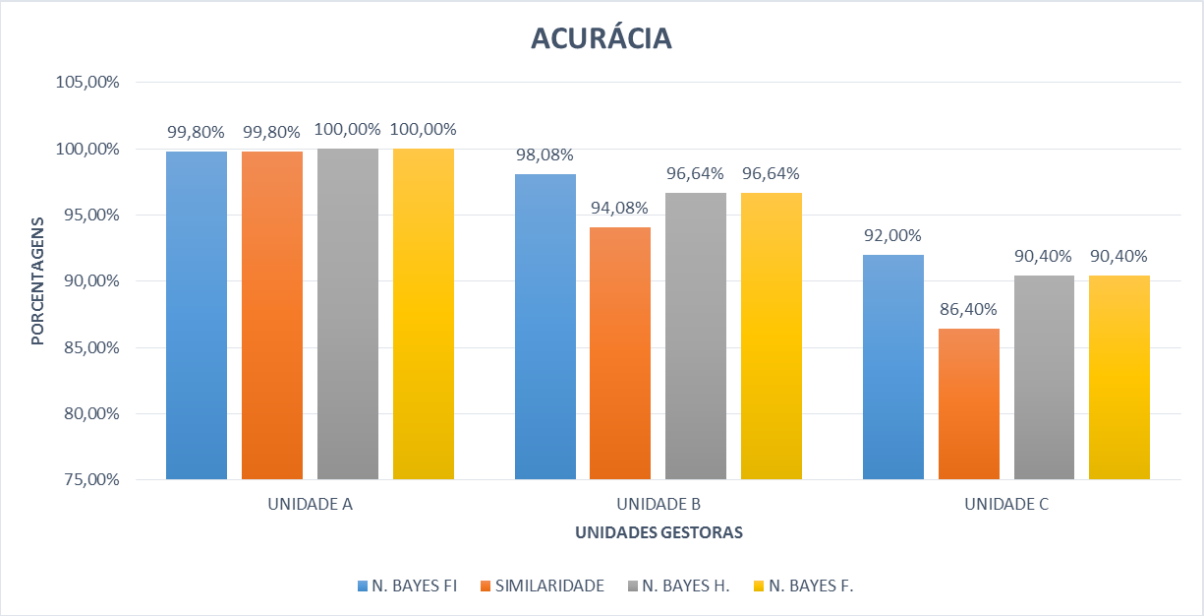

Gráfico 01. Gráfico da métrica Acurácia.

No Gráfico 02, é notável o bom desempenho do Naïve Bayes – Frequência Inversa na maioria das unidades, apesar do péssimo desempenho na Unidade B. Em média, Naïve Bayes – Híbrido e Naïve Bayes – Frequência foram melhores do que Naïve Bayes – Frequência Inversa. Mesmo assim, Naïve Bayes – Frequência Inversa é a melhor abordagem em termos de Precisão, pois, como foi dito anteriormente, classificou erroneamente um número inferior de sentenças em relação aos demais algoritmos.

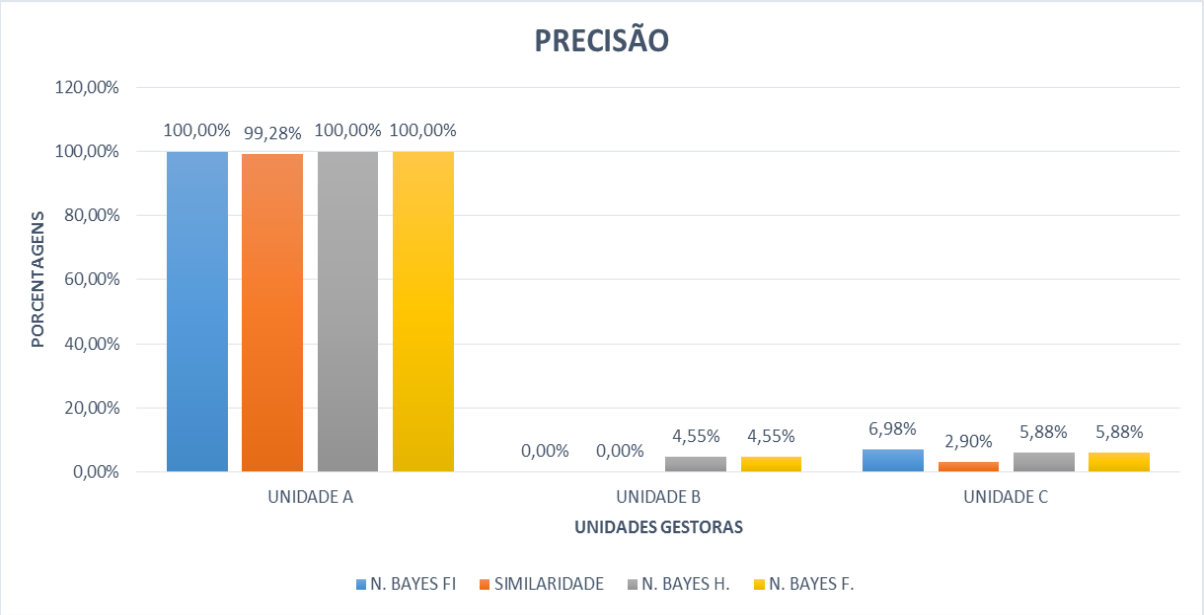

Gráfico 02. Gráfico da métrica Precisão.

Observando o Gráfico 03, concluímos o ótimo desempenho de Naïve Bayes – Híbrido e Naïve Bayes – Frequência, possuindo 100 % em todas as unidades. Assim como é perceptível o baixo desempenho de Similaridade na métrica Cobertura. Portanto, Naïve Bayes – Híbrido e Naïve Bayes – Frequência, por possuírem as melhores porcentagens de Cobertura, são os melhores algoritmos em termos de Cobertura.

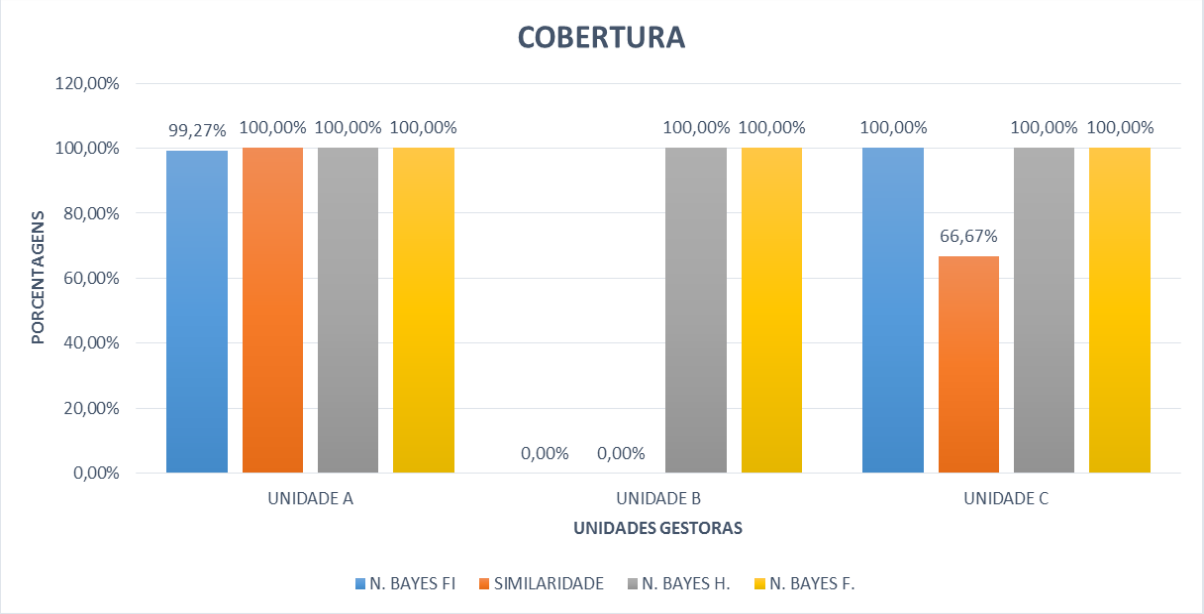

Gráfico 03. Gráfico da métrica Cobertura.

Analisando o Gráfico 04, é notável a qualidade de Naïve Bayes – Híbrido e Naïve Bayes – Frequência, apesar de possuir a média harmônica um pouco inferior à de Naïve Bayes – Frequência Inversa na Unidade C. Já Similaridade possui, em média, a menor porcentagem de medida F. Assim, Naïve Bayes – Híbrido e Naïve Bayes – Frequência possuem, em média, as melhores porcentagens de Medida F, consequentemente, são os melhores algoritmos em termos de Medida F.

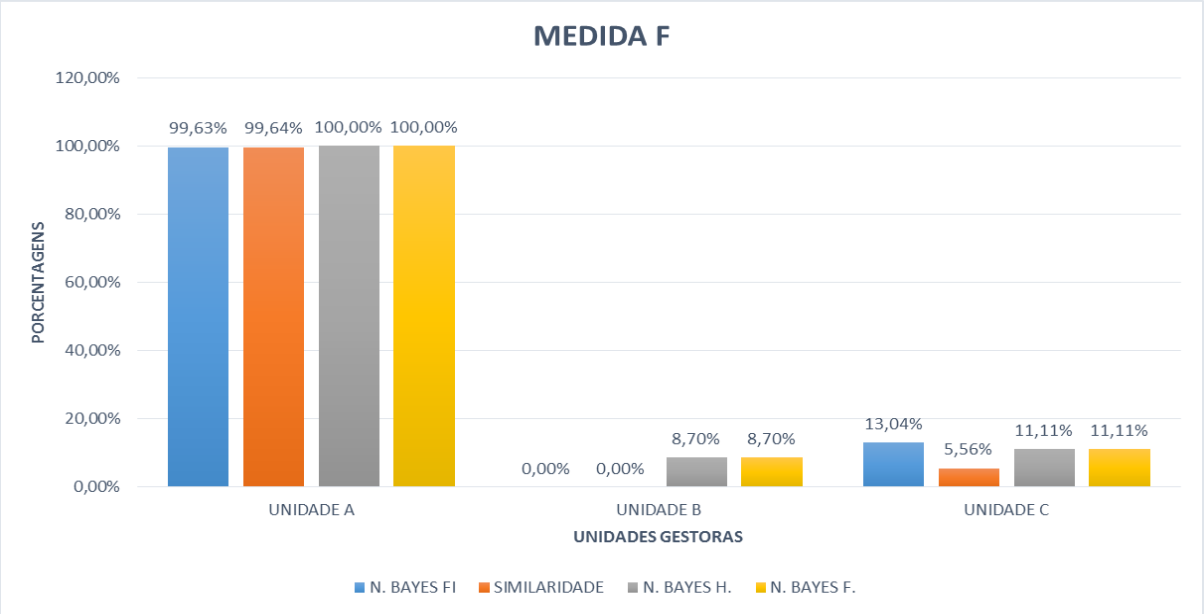

Gráfico 04. Gráfico da métrica Medida F.

Observando o Gráfico 05, é evidente o ótimo desempenho do algoritmo de Similaridade por possuir os menores tempos de execução em todas as unidades gestoras. Assim, Similaridade é a melhor abordagem em termos de Tempo Médio de Execução.

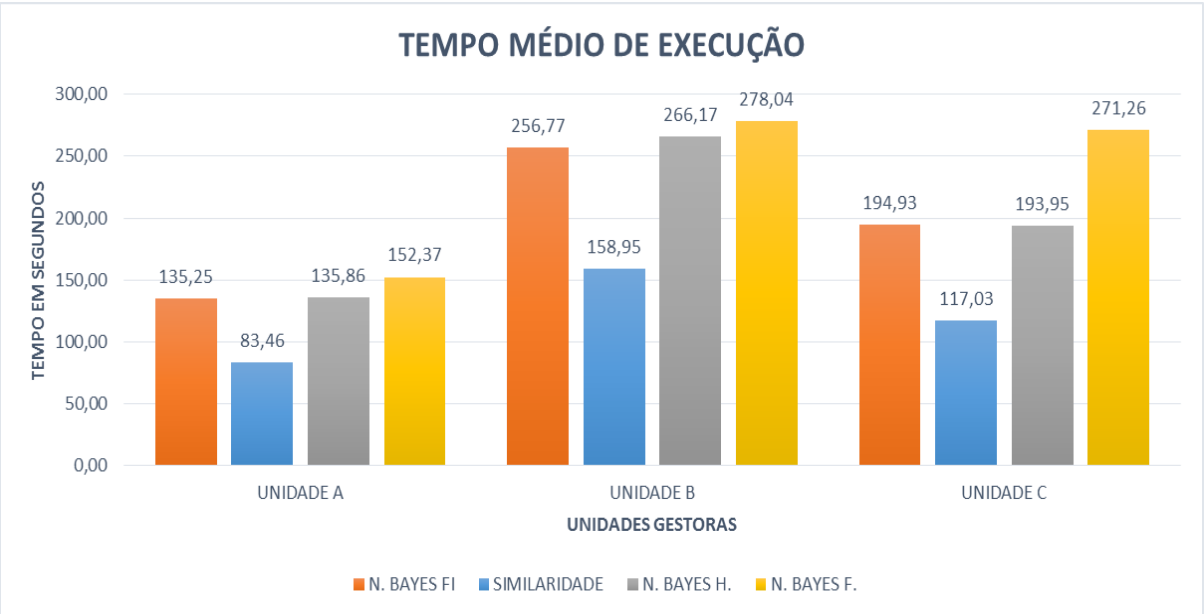

Gráfico 05. Gráfico da métrica Tempo Médio de Execução.

Na tentativa de descobrir o motivo de todos algoritmos terem classificados erroneamente uma quantidade considerável de evidências nas unidades B e C, foram analisadas as classificações de duas conformidades que foram consideradas como evidências por todas as abordagens: “INAUGURACAO DE CONJUNTO HABITACIONAL,CLINICA DE SAUDE E CEN” (Unidade B) e “LEVAR PACIENTES P/REALIZACAO DE EXAMES” (Unidade C). Para classificar a primeira sentença, foi considerado apenas um termo cujo radical é “saud”, sendo 06 e 09 as quantidades de amostras “Em Conformidade” e “Evidência”, respectivamente. Já para a segunda sentença, foram considerados dois termos de radicais “pacient” e “exam”, sendo 01 a quantidade de “Em Conformidade” para o radical “pacient” e 06 a quantidade de “Evidência” para o radical “exam”. Portanto, o processo de *Stemming* influenciou na classificação errônea dessas sentenças, podendo até ter influenciado nas demais sentenças classificadas de forma errada.

Finalizando a análise, na maioria das métricas e unidades analisadas, conclui-se que Naïve Bayes – Frequência Inversa, para o contexto abordado neste trabalho, foi o algoritmo que obteve melhor desempenho e qualidade para classificar sentenças, consequentemente, possibilitando melhores resultados para apoiar a decisão dos auditores na detecção de irregularidades no pagamento de diárias.

6 CONCLUSÃO

A principal contribuição deste trabalho foi a avaliação dos algoritmos de mineração de texto disponíveis na ferramenta TextMining em termos de desempenho e qualidade para detectar irregularidades em históricos de contas públicas. O trabalho foi consolidado pela realização de um estudo de caso, o qual, a partir das unidades escolhidas, determinou Naïve Bayes – Frequência Inversa como a melhor abordagem para identificação de evidências. De posse do melhor algoritmo, esse pode ser utilizado para tornar mais efetivo o trabalho do auditor na identificação de irregularidades, auxiliando-o na tomada de decisão.

O referido trabalho possibilitou um melhor entendimento do processo de KDT e da avaliação de desempenho e qualidade de classificadores, bem como despertou o interesse por conhecimentos das áreas de Processamento de Linguagem Natural (PLN) e Recuperação de Informação (RI) com o intuito de melhorar o processo de KDT.

Foi desenvolvido um novo algoritmo, Similaridade, aproveitando as rotinas de pré-processamento para calcular a similaridade entre uma amostra e uma sentença a ser classificada.

Por meio do estudo de caso, foi constatado que não houve diferença no desempenho e qualidade dos algoritmos Naïve Bayes – Híbrido e Naïve Bayes – Frequência, bem como a possibilidade do processo de radicalização ter influenciado na classificação errônea de evidências.

6.1 Trabalhos Futuros

Como consequência deste trabalho, é possível vislumbrar os possíveis trabalhos futuros:

- Adição de novas funcionalidades no módulo de dicionário como, por exemplo, a submissão de um arquivo contendo amostras relevantes, possibilitando maior agilidade na criação do modelo de conhecimento.
- Implementar novos algoritmos de classificação como, por exemplo, Redes Neurais, *SVM*, Árvores de Decisão, *K-means*, *K-Nearest-Neighbor*, entre outros. Assim como realizar uma nova avaliação de desempenho e qualidade para cada novo algoritmo implementado.
- Implementar novas funções de similaridade e avaliar qual função possibilita melhores resultados na classificação de evidências.

REFERÊNCIAS

AMOOEE, G.; MINAEI-BIDGOLI, B.; BAGHERI-DEHNAVI, M. **A Comparison Between Data Mining Prediction Algorithms for Fault Detection (Case study: Ahanpishegan co.)**. 2011. Disponível em: <http://ijcsi.org/papers/IJCSI-8-6-3-425-431.pdf>. Acesso em: 05 de out. 2014.

BALINSKI, R. **Filtragem de Informações no Ambiente do Direito**. Dissertação (Mestre em Informática). Porto Alegre: PPGC da UFRGS, 2002. 87 p.

BHANDARI, I.; COLET, E.; PARKER, J.; PINES, Z.; PRATAP, R.; RAMANUJAM, K. **Brief Application Description Advanced Scout: Data Mining and Knowledge Discovery in NBA Data**. 1997. Disponível em: <http://download.springer.com/static/pdf/801/art%253A10.1023%252FA%253A1009782106822.pdf?auth66=1411048266_0fcad66d2a458fae6cb5b784d231a58b&ext=.pdf>. Acesso em: 20 de ago. 2014.

BOA SAÚDE. **Exames de Rotina**. Disponível em: <http://www.boasaude.com.br/exames-de-rotina/todos/pagina/1/>. Acesso em: 20 de nov. 2014.

BRAMER, M. **Principles of Data Mining**. New York: Springer London, 2007.

BRILHADORI, M; LAURETTO, M. S. **Estudo comparativo entre algoritmos de árvores de classificação e máquinas de vetores suporte, baseados em ensembles de classificadores**. 2013. Disponível em: < http://www.lbd.dcc.ufmg.br/colecoes/sbsi/2013/009.pdf>. Acesso em: 10 de nov. 2014.

CASTRO, D. P. de. **Auditoria e controle interno na administração pública: evolução do controle interno no Brasil: do Código de Contabilidade de 1992 até a criação da CGU em 2003: guia para atuação das auditorias e organização dos controles internos nos Estados, municípios e ONGs**. 2ª ed. São Paulo: Atlas, 2009.

COLAÇO JÚNIOR, M. **Projetando Sistemas de Apoio à Decisão Baseados em Data Warehouse**. Rio de Janeiro: Axcel Books, 2004.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. **From Data Mining to Knowledge Discovery in Databases**. 1996. Disponível em: <http://www.aaai.org/ojs/index.php/aimagazine/article/view/1230>. Acesso em: 15 de mai. 2014.

FELDMAN, R.; DAGAN, I. **Knowledge Discovery in Textual Databases (KDT)**. 1995. Disponível em: <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.47.7462&rep=rep1&type=pdf>. Acesso em: 15 de mai. 2014.

GHAZVINI, A.; AWWALU, J.; BAKAR, A. A. **Comparative Analysis of Algorithms in Supervised Classification: A Case study of Bank Notes Dataset**. 2014. Disponível em: <http://www.ijcttjournal.org/Volume17/number-1/IJCTT-V17P109.pdf>. Acesso em: 10 de out. 2014.

GUIMARÃES, D. T. **Dicionário de Termos Médicos e de Enfermagem**. São Paulo: Rideel, 2002.

GONZALEZ, M.; LIMA, V. L. S. **Recuperação de Informação e Processamento da Linguagem Natural**. In: XXIII Congresso da Sociedade Brasileira de Computação. Anais da III Jornada de Mini-Cursos de Inteligência Artificial. Campinas: [s.n.], v. III, 2003. p. 347-395.

HAN, J.; KAMBER, M.; PEI, J. **Data Mining: Concepts and Techniques**. 3ª ed. San Francisco: Morgan Kaufmann Publishers, 2011.

LAMKANFI, A.; DEMEYER, S.; SOETENS, Q. D.; VERDONCK, T. **Comparing Mining Algorithms for Predicting the Severity of a Reported Bug**. 2011. Disponível em: < http://ieeexplore.ieee.org/ielx5/5740650/5741244/05741332.pdf?tp=&arnumber=5741332&isnumber=5741244>. Acesso em: 30 de nov. 2014.

MAGALHÃES, C. C. **MinerJur: Uma ferramenta para mineração de bases de jurisprudência**. Dissertação (Mestrado em Sistemas e Computação). Salvador: Universidade Salvador, 2008. 144 p.

MCCALLUM, A.; NIGAM, K. **A Comparison of Event Models for Naive Bayes Text Classification**. 1998. Disponível em: <http://citeseerx.ist.psu.edu/viewdoc/download;jsessionid=6D492C0CABE07EEE0E3BF2DCD8DC1628?doi=10.1.1.46.1529&rep=rep1&type=pdf>. Acesso em: 05 de nov. 2014

MORAIS, E. A. M.; AMBRÓSIO, A. P. L. **Mineração de Texto**. Relatório Técnico. Instituto de Informática da Universidade Federal de Goiás, 2007. 30p.

ORENGO, V. M.; HUYCK, C. **A Stemming Algorithm for the Portuguese Language**. 2001. Disponível em: <http://homes.dcc.ufba.br/~dclaro/download/mate04/Artigo%20Erick.pdf>. Acesso em: 06 de jun. 2014.

PINHO, R. C. de S. **Fundamentos de auditoria: auditoria contábil: outras aplicações de auditoria**. São Paulo: Atlas, 2007.

SÁ, H. R. de. **Seleção de Características para Classificação de Texto**. Recife: UFPE, 2008. 57 p.

SERGIPE. **Decreto Nº. 12.424, de 12 de setembro de 1991**. Regulamenta a concessão de diária aos servidores civis da Administração Estadual Direta, do Poder Executivo, que se deslocarem para localidades situadas dentro ou fora do Estado de Sergipe. Controladoria-Geral do Estado de Sergipe.

SILVA, R. C. L. da. **Compacto Dicionário Ilustrado de Saúde**. 2ª ed. São Caetano do Sul: Yendis Editora, 2007.

SOARES, A. M. **A Mineração de Texto na Análise de Contas Públicas Municipais**. Dissertação (Mestrado Profissional em Computação Aplicada). Fortaleza: Universidade Estadual do Ceará, 2010. 85 p.

SOUZA, E. N. P. de; CLARO, D. B. **Detecção Multilíngue de Serviços Web Duplicados Baseada na Similaridade Textual**. 2014. Disponível em: <http://www.lbd.dcc.ufmg.br/colecoes/sbsi/2014/0043.pdf>. Acesso em: 07 de jun. 2014.

SOUZA, J. G. de. **Uma aplicação de Mineração de Texto para Descoberta de Características Psicológicas de Indivíduos**. Itabaiana: UFS, 2011. 62 p.

TING, S. L.; IP, W. H.; TSANG, A. H. C. **Is Naïve Bayes a Good Classifier for Document Classification?**. 2011. Disponível em: < http://www.sersc.org/journals/IJSEIA/vol5_no3_2011/4.pdf>. Acesso em: 30 de nov. 2014.

TRIBUNAL DE CONTAS DE SERGIPE. **SISAP**. Disponível em: <http://www.tce.se.gov.br/sitev2/sisap.php>. Acesso em: 25 de nov. 2014.

VIJAYARANI, S.; MUTHULAKSHMI, S. **Comparative Analysis of Bayes and Lazy Classification Algorithms**. 2013. Disponível em: <http://www.ijarcce.com/upload/2013/august/34-h-Uma%20Gopalakrishnan%20-Comparative%20Analysis%20of%20Bayes%20and%20Lazy%20classification%20algorithms.pdf>. Acesso em: 15 de out. 2014.

WEISS, S. M.; INDURKHYA, N.; ZHANG, T. **Fundamentals of Predictive Text Mining**. New York: Springer London, 2010.

WITTEN, I. H.; FRANK, E. **Data Mining: Practical Machine Learning Tools and Techniques**. 2ª ed. San Francisco: Elsevier, 2005.

WIVES, L. K. **Tecnologias de Descoberta de Conhecimento em Textos Aplicadas à Inteligência Competitiva**. Exame de Qualificação EQ-069 (Doutorado). Porto Alegre: PPGC da UFRGS, 2002. 116 p.

WIKCIONÁRIO. **Medicina**. Disponível em: <http://pt.wiktionary.org/wiki/medicina>. Acesso em: 20 de nov. 2014.

WIKIPÉDIA. **Lista de Doenças**. Disponível em: <http://pt.wikipedia.org/wiki/Lista_de_doen%C3%A7as>. Acesso em: 20 de nov. 2014.

WIKIPÉDIA. **Lista de doenças causadas por seres vivos**. Disponível em: <http://pt.wikipedia.org/wiki/Lista_de_doen%C3%A7as_causadas_por_seres_vivos>. Acesso em: 20 de nov. 2014.

APÊNDICE

APÊNDICE A – Pseudocódigo do algoritmo implementado

```
Registro Ocorrencia
        steam : caractere
        frequencia : inteiro
        tf : real
Fim_Registro

Registro ResultadoGenerico
        indiceAmostra : inteiro
        classe : booleano
        valor : real
Fim_Registro

Registro TermoAmostra
        indiceAmostra : inteiro
        tf_idf : real
        classe : booleano
Fim_Registro

função VerificaSentenca(linha : caractere, dicionario : Dicionario) : ResultadoGenerico
        Declare
                NBayes : NaiveBayes // Objeto NBayes para aproveitar a função CalcularOcorrenciaTermos
                ga : GerenciadorAmostra // Objeto para utilizar funções para gerenciar amostras do dicionário
                gt : GerenciadorTermo // Objeto para gerenciar termos das amostras
                contSteam, contAmostra, indAmostra : inteiro
                fimListaScores, posicaoAtual, quantVerdadeiro, quantFalso : inteiro
                ocorrencias : Lista de Ocorrencia
                scores : Lista de ResultadoGenerico
                amostras : Lista de TermoAmostra
                resultado : ResultadoGenerico

        linha <- TransformaCaracteresEmMinusculo(linha)
        linha <- RemoveEspacoEmBrancoADireitaAEsquerda(linha)
        ocorrencias <- NBayes.CalcularOcorrenciaTermos(linha)

        // Determinando os scores para as amostras
        para contSteam = 1 até ocorrencias.QuantidadeElementos() faça


                // Verifica se existe o termo
                se ( gt.ExisteTermo(ocorrencias[contSteam].steam) ) então

                        // Verifica se existe amostras contendo o termo para o dicionário passado
                        se ( gt.ExisteAmostra(ocorrencias[contSteam].steam, dicionario) ) então

                                amostras <- gt.ObterTermoAmostra(ocorrencias[contSteam].steam, dicionario)

                                para contAmostra = 1 até amostras.QuantidadeElementos() faça

                                        se ( scores.PossuiElementoPorIndiceAmostra(amostras[contAmostra].indiceAmostra) )
                                                // Obter a posição do elemento que possui indiceAmostra na lista scores
                                                indAmostra <- scores.RetornaPosicaoElementoPorIndiceAmostra(amostras[contAmostra].indiceAmostra)

                                                // Atualizando o score da amostra
                                                scores[indAmostra].valor <- scores[indAmostra].valor + ( amostras[contAmostra].tf_idf * ocorrencias[contSteam].tf )

                                        senão
                                                // Inicializando resultado
                                                resultado.indiceAmostra <- amostras[contAmostra].indiceAmostra
                                                resultado.valor <- amostras[contAmostra].tf_idf * ocorrencias[contSteam].tf
                                                resultado.classe <- ga.ObterClasseAmostraPorIndice(amostras[contAmostra].indiceAmostra)

                                                // Adicionando resultado na lista scores
                                                scores.AdicionarElemento(resultado)
                                        fim-se
                                fim-para
                        fim-se
                fim-se
        fim-para

        // Determinando a classe da sentença
        fimListaScores <- scores.QuantidadeElementos()

        se ( fimListaScores > 1 ) então
```

```
                // Ordenando o vetor de scores em ordem crescente
                scores.OrdenarLista()

                se ( scores[fimListaScores].valor > scores[fimListaScores - 1].valor ) então
                        // Retorna o objeto com maior score
                        retorne scores[fimListaScores]
                senão
                        // Para os mesmos scores, retorna a maior ocorrência do label
                        posicaoAtual <- fimListaScores - 1
                        quantVerdadeiro <- 0
                        quantFalso <- 0

                        se ( scores[fimListaScores].classe = verdadeiro ) então
                                quantVerdadeiro <- quantVerdadeiro + 1
                        senão
                                quantFalso <- quantFalso + 1
                        fim-se

                        enquanto ( posicaoAtual >= 0 ) faça
                                se ( (posicaoAtual >= 1) E (scores[fimListaScores].valor = scores[posicaoAtual].valor) ) então
                                        se ( scores[posicaoAtual].classe = verdadeiro ) então
                                                quantVerdadeiro <- quantVerdadeiro + 1
                                        senão
                                                quantFalso <- quantFalso + 1
                                        fim-se
                                        posicaoAtual <- posicaoAtual - 1
                                senão
                                        se ( quantVerdadeiro > quantFalso ) então
                                                // Inicializando resultado
                                                resultado.indiceAmostra <- 0
                                                resultado.valor <- scores[fimListaScores].valor
                                                resultado.classe <- verdadeiro

                                                // Quando é classificado como Evidência
                                                retorne resultado

                                        senão
                                                se ( quantFalso > quantVerdadeiro ) então
                                                        // Inicializando resultado
                                                        resultado.indiceAmostra <- 0
                                                        resultado.valor <- scores[fimListaScores].valor
                                                        resultado.classe <- falso

                                                        // Quando é classificado como Em Conformidade
                                                        retorne resultado

                                                senão
                                                        fimListaScores <- posicaoAtual
                                                        posicaoAtual <- posicaoAtual - 1
                                                        quantVerdadeiro <- 0
                                                        quantFalso <- 0
                                                        se ( fimListaScores >= 1 ) então
                                                                se ( scores[fimListaScores].classe = verdadeiro ) então
                                                                        quantVerdadeiro <- quantVerdadeiro + 1
                                                                senão
                                                                        quantFalso <- quantFalso + 1
                                                                fim-se
                                                        fim-se
                                                fim-se
                                        fim-se
                                fim-se
                        fim-enquanto

                        // Inicializando resultado
                        resultado.indiceAmostra <- 0
                        resultado.valor <- 0.0
                        resultado.classe <- falso

                        // Quando não classifica
                        retorne resultado
                fim-se
        senão

                se ( fimListaScores == 1 ) então
                        // Retorna o único score
                        retorne scores[fimListaScores]
                senão
                        // Inicializando resultado
                        resultado.indiceAmostra <- 0
                        resultado.valor <- 0.0
                        resultado.classe <- falso

                        // Quando não classifica
                        retorne resultado
                fim-se
        fim-se
fim-função
```

APÊNDICE B – Lista de termos mais comuns e relevantes na área da saúde por categoria

1. PROCEDIMENTOS, TRATAMENTOS E CIRURGIAS MÉDICAS: ABLATIVA; ABORTO; ACUPUNTURA; ALOPATIA; AMPUTAÇÃO; ANESTESIA; ANTI-SEPSIA; APENDICECTOMIA; ASSEPSIA; AUSCULTAÇÃO; AUTÓPSIA; BARIÁTRICA; BIÓPSIA; CABEÇA; CARDÍACA; CAUTERIZAÇÃO; CHECK-UP; CIRURGIA; COLUNA; COSTURA; DRENO; ELETROCIRURGIA; EXAME; HEMODIÁLISE; HERNIOTOMIA; HIDROTERAPIA; HISTERECTOMIA; HOMEOPATIA; IMPLANTE; LAPAROSCOPIA; LAVAGEM; LIPOASPIRAÇÃO; LOBOTOMIA; MASSAGEM; MASTECTOMIA; NEFRECTOMIA; NEUROCIRURGIA; OBTURAÇÃO; OCLUSÃO; OCUPACIONAL; OPERAÇÃO; OPERAÇÃO CESARIANA OU CESÁREA; ORTOPÉDICA; PESCOÇO; PLÁSTICA; PROFILÁTICA; PSICANÁLISE; PUNÇÃO; QUIMIOTERAPIA; QUIROPATIA; QUIROPRAXIA; RADIOCIRURGIA; RADIOSCOPIA; RADIOTERAPIA; RINOTOMIA; SONOTERAPIA; SOROTERAPIA; TERAPIA; TRANSFUSÃO; TRANSFUSÃO DE SANGUE; TRANSPLANTAÇÃO; TRANSPLANTE; TRAQUEOPLASTIA; TRATAMENTO; TRATAMENTO DE CHOQUE; TREPANAÇÃO; ULTRA-SONOCIRURGIA; VACINAÇÃO; VASECTOMIA; VIDEOCIRURGIA; ZONULOTOMIA.

2. PROFISSIONAIS DA SAÚDE: ANDROLOGISTA; ANESTESIOLOGISTA; ANESTESISTA; CARDIOLOGISTA; CARDIÓLOGO; CIRURGIÃO; CIRURGIÃO-DENTISTA; DENTISTA; DERMATOLOGISTA; DOUTOR; ENDOCRINOLOGISTA; ENDÓCRINO; ENFERMEIRA; EPIDEMIOLOGISTA; FISIOTERAPEUTA; FONOAUDIÓLOGO; GASTRENTEROLOGISTA; GERIATRA; GERONTOLOGISTA; GERONTÓLOGO; GINECOLOGISTA; HOMEOPATA; IMUNOLOGISTA; LEGISTA; MÉDICO; MÉDICO-LEGISTA; NEFROLOGISTA; NEFRÓLOGO; NEONATOLOGISTA; NEUROCIRURGIÃO; NEUROLOGISTA; NUTRICIONISTA; OBSTETRA; OFTALMOLOGISTA; OFTALMÓLOGO; ONCOLOGISTA; OPTOMETRISTA; ORTOPEDISTA; OSTEOPATA; OTORRINOLARINGOLOGISTA; PATOLOGISTA; PEDIATRA; PODÓLOGO; PROCTOLOGISTA; PSICANALISTA; PSICOTERAPEUTA; PSICÓLOGO; PSIQUIATRA; RADIOLOGISTA; REUMATOLOGISTA; SANITARISTA;

SEROLOGISTA; TERAPEUTA; TERAPISTA; TRAUMATOLOGISTA; UROLOGISTA; URÓLOGO.

3. CIÊNCIAS DA SAÚDE: ANDROLOGIA; ANESTESIOLOGIA; ANGIOLOGIA; AUXOLOGIA; BIOMEDICINA; CANCEROLOGIA; CARDIOLOGIA; DERMATOLOGIA; ENDOCRINOLOGIA; EPIDEMIOLOGIA; ESTOMATOLOGIA; FISIOTERAPIA; FONIATRIA; FONOAUDIOLOGIA; GASTRENTEROLOGIA; GERIATRIA; GERONTOLOGIA; GINECOLOGIA; IMUNOLOGIA; MASTOLOGIA; NEFROLOGIA; NEONATOLOGIA; NEUROLOGIA; NEURORRADIOLOGIA; NUTRIÇÃO; OBSTETRÍCIA; ODONTOLOGIA; OFTALMOLOGIA; ONCOLOGIA; OPTOMETRIA; ORTODONTIA; ORTOPEDIA; OTORRINOLARINGOLOGIA; PATOLOGIA; PEDIATRIA; PNEUMOLOGIA; PODOLOGIA; PROCTOLOGIA; PSICOLOGIA; PSICOTERAPIA; PSIQUIATRIA; RADIOLOGIA; REUMATOLOGIA; SEROLOGIA; SINTOMATOLOGIA; SOMATOLOGIA; TERAPÊUTICA; TRAUMATOLOGIA; UROLOGIA.

4. PROBLEMAS MÉDICOS: ABSCESSO; ALERGIA; APENDICITE; ASFIXIA; BACTÉRIA; CRISE; CÁLCULO RENAL; DERRAME; DISFAGIA; DISFUNÇÃO; DISTENSÃO; DISTROFIA; DOENÇA; EDEMA; ENXAQUECA; ESCORIAÇÃO; ESPASMO; ESTIRAMENTO; FRATURA; FUNGO; HANSENÍASE; HEMORRAGIA; INFARTO; INFECÇÃO; INFLAMAÇÃO; INSOLAÇÃO; INSUFICIÊNCIA; INSÔNIA; LESÃO; MUDEZ; OBESIDADE; PEDRA NO RIM; PNEUMONIA; PROTOZOÁRIO; REAÇÃO; RECORRÊNCIA; REJEIÇÃO; REUMATISMO; SEQUELA; SINTOMA; STRESS; SUFOCAMENTO; SUFOCAÇÃO; SÍNCOPE; SÍNDROME; TORCICOLO; TRAUMA; TRAUMATISMO; VERME; VERMINOSE; VÍRUS.

5. ADJETIVOS DA ÁREA DA SAÚDE: CIRÚRGICO; CITOPATOLÓGICO; CLÍNICO; DERMATOLÓGICO; ECOCARDIOGRÁFICO; ELETRENCEFALOGRÁFICO; EPIDEMIOLÓGICO; FISIOLÓGICO; FISIOTERÁPICO; FONOAUDIOLÓGICO; FÍSICO; GERIÁTRICO; GERONTOLÓGICO; GINECOLÓGICO; HIPOCRÁTICO; HOMEOPÁTICO; HOSPITALAR; IDIOSSINCRÁTICO; IMUNITÁRIO; IMUNOLÓGICO; LABORATORIAL; MEDICINAL; MÉDICO-HOSPITALAR; NEFROLÓGICO;

NEUROLÓGICO; OFTALMOLÓGICO; ONCOLÓGICO; OPERATÓRIO; OPTOMÉTRICO; ORTOPÉDICO; PARALÍTICO; PARAMÉDICO; PATOLÓGICO; POLICLÍNICA; PROCTOLÓGICO; PROFILÁTICO; PSIQUIÁTRICO; QUADRIPLÉGICO; QUIMIOTERÁPICO; QUIROPRÁTICO; RADIOGRÁFICO; RADIOLÓGICO; RADIOSCÓPICO; REUMATOLÓGICO; SEROLÓGICO; SINTOMATOLÓGICO; SINTOMÁTICO; SOMATOLÓGICO; TERAPÊUTICO; TRAUMATOLÓGICO; TRAUMÁTICO; UROLÓGICO.

6. TERMOS GERAIS: AMBULATÓRIO; ASSISTÊNCIA MÉDICA; ATENDIMENTO MÉDICO; CARDIOGRAMA; CLÍNICA; CONSULTA MÉDICA; CONSULTÓRIO; DIAGNÓSTICO MÉDICO; DESPESA MÉDICA; ELETROENCEFALOGRAMA; EMERGÊNCIA; EXAMES MÉDICOS; HISTÓRICO DE SAÚDE; HOSPITAL; LAUDO; PERÍCIA MÉDICA; POLICLÍNICA; PROCEDIMENTO DE SAÚDE; PRONTO-SOCORRO; QUADRO CLÍNICO; RADIOGRAFIA; REVISÃO MÉDICA; VACINA.

7. DOENÇAS: ACNE; ANEMIA; ANSIEDADE; ARTRITE; ARTROSE; ASCARIDÍASE; ASMA; BERIBÉRI; CANCRO, TUMOR OU CÂNCER; CIRROSE HEPÁTICA; CÁRIE; DENGUE; DEPRESSÃO; DERMATITE SEBORRÉICA, SEBORRÉIA OU CASPA; DERMATOFITOSE, MICOSE; DIABETES INSIPIDUS MELLITUS; DIARREIA; DISENTERIA AMÉBICA OU AMEBIANA, AMEBÍASE; DISENTERIA BACTERIANA OU SHIGELOSE; DISLIPIDEMIA; DOENÇA DE CHAGAS, CHAGUISMO OU TRIPANOSSOMÍASE AMERICANA; ÉBOLA; EPILEPSIA; ESCABIOSE OU SARNA; ESCLEROSE MÚLTIPLA; ESOFAGITE; ESQUISTOSSOMOSE OU BILHARZÍASE; FARINGITE; FEBRE; FIBROSE CÍSTICA; GASTRITE; GIARDIOSE OU GIARDÍASE; GLAUCOMA; GOTA; HEPATITE; HERPES; HIPERCOLESTEROLEMIA; HIPERPARATIROIDISMO; HIPERTENSÃO ARTERIAL OU PULMONAR; HIPERTIROIDISMO; HIPOTIROIDISMO; LEISHMANIOSE, LEISHMANÍASE, CALAZAR OU ÚLCERA DE BAURU; LEUCEMIA MIELOIDE AGUDA; LEUCEMIA OU LINFOMA DE CÉLULAS T DO ADULTO; LÚPUS ERITEMATOSO SISTÊMICO; MAL DE ALZHEIMER; MAL DE PARKINSON; MALÁRIA OU PALUDISMO; MENINGITE; OSTEOPOROSE; PNEUMONIA; RUBÉOLA OU RUBELA; SARAMPO; SINUSITE; SÍFILIS; SÍNDROME DA IMUNODEFICIÊNCIA ADQUIRIDA, AIDS OU SIDA HIV;

TRANSTORNOS ALIMENTARES; TUBERCULOSE; ÚLCERA; VARICELA OU CATAPORA.

8. EXAMES: 25-HIDROXIVITAMINA D OU 25(OH)D; ACIDO ÚRICO NO SANGUE; ALBUMINA; ALTERAÇÕES DO FERRO E DE SUA CAPACIDADE DE FIXAÇÃO; AUDIOMETRIA VON BEKESY; BILIRRUBINA NA URINA; BILIRRUBINA NO SANGUE (DIRETA, INDIRETA E TOTAL). PROVAS DE FUNÇÃO HEPÁTICA (BILIRRUBINAS, ELETROFORESE DE PROTEÍNAS. FA, TGO, TGP E GAMA-PGT); CÁLCIO NO SANGUE; CARDIOLIPINA, AUTO-ANTICORPOS IGG; CITOGENÉTICA DIAGNÓSTICO PRÉ-NATAL; CLEARANCE DE URÉIA; CLEARANCE DE CREATININA; CLORO NO SANGUE (CL); COLESTEROL TOTAL; COLONOSCOPIA; CORTISOL PLASMÁTICO; CREATININA NO SANGUE; CREATINOFOSFOQUINASE OU CPK; CULTURA BACTERIOLÓGICA DO SANGUE (HEMOCULTURA); CURVA DE TOLERÂNCIA A GLICOSE; DENGUE, SOROLOGIA; DENSITOMETRIA ÓSSEA; DIAGNÓSTICO LABORATORIAL DA HEPATITE; DOPPLER SCAN COLORIDO ARTERIAL DE MEMBRO INFERIOR E SUPERIOR, COLORIDO DE VÍSCERAS ABDOMINAIS, DE CARÓTIDAS E VERTEBRAIS, VENOSO DE MEMBRO INFERIOR - UNILATERAL; ECOCARDIOGRAFIA, ECODOPPLERCARDIOGRAMA TRANSTORÁCICO; ELETROCARDIOGRAMA (ECG); ENDOSCOPIA; EXAME PARASITOLÓGICO DE FEZES; FERRITINA NO SANGUE; FERRO SÉRICO; FIBRINOGÊNIO PLASMÁTICO; FIBROSE CÍSTICA, ESTUDO GENÉTICO; FOSFATASE ÁCIDA, ALCALINA, ALCALINA NEUTROFÍLICA OU LEUCOCITÁRIA; GAMA-GLUTAMIL TRANSFERASE (GGT); GLICEMIA PÓS PRANDIAL; HIV AIDS (SÍNDROME DE IMUNO DEFICIÊNCIA ADQUIRIDA) (EXAME DE WESTERN BLOT E TESTE DE ELISA); HPV CAPTURA HÍBRIDA PROCEDIMENTO DIAGNÓSTICO POR CAPTURA HÍBRIDA; HEMOGRAMA COM CONTAGEM DE PLAQUETAS OU FRAÇÕES (ERITROGRAMA, ERITRÓCITOS, LEUCÓCITOS, LEUCOGRAMA, PLAQUETAS); HEMOSSEDIMENTAÇÃO; HORMÔNIO DE CRESCIMENTO NO SANGUE. HORMÔNIO SOMATOTRÓFICO (STH); HORMÔNIO LUTEINIZANTE NO PLASMA; HORMÔNIO PARATIREOIDEANO NO SANGUE; IMUNOGLOBULINAS E TOTAL, G, A E M NO SANGUE; INSULINA NO SANGUE; MAGNÉSIO NO SANGUE (MG+); MAMOGRAFIA; MICROALBUMINÚRIA; PAPANICOLAU (CITOLOGIA

VAGINAL); PEPTÍDEO C; POTÁSSIO NO SANGUE (K+); PROTEÍNA C REATIVA; RAIO X DA PERNA, DO ANTEBRAÇO, DO BRAÇO, DOS SEIOS DA FACE; RESSONÂNCIA MAGNÉTICA (RM) DE CRÂNIO (ENCÉFALO), DA COLUNA; SANGUE OCULTO NAS FEZES, PESQUISA; TSH; TEMPO DE COAGULAÇÃO E DE RETRAÇÃO DO COÁGULO; TESTE ERGOMÉTRICO; TESTOSTERONA LIVRE; TIROXINA (T4); TOMOGRAFIA COMPUTADORIZADA (TC) DE ABDOMEM, DE COLUNA VERTEBRAL, DE CRÂNIO, DE TÓRAX, DOS SEIOS PARANASAIS; TRANSAMINASE OXALACÉTICA (TGO), PIRÚVICA (TGP); TRANSFERRINA; TRI IODO TIRONINA (T3); TRIGLICÉRIDES; ULTRASSONOGRAFIA, ULTRA-SONOGRAFIA (US), ULTRASSOM ABDOMINAL ABDOME INFERIOR MASCULINO OBSTÉTRICA (BEXIGA, PRÓSTATA E VESÍCULAS SEMINAIS) ABDOME INFERIOR FEMININO (BEXIGA, ÚTERO, OVÁRIO E ANEXOS) ABDOME TOTAL (INCLUI PELVE) ABDOME SUPERIOR (FÍGADO, VIAS BILIARES, VESÍCULA, PÂNCREAS, BAÇO), DA TIREÓIDE, DA MAMA; URINA (ANÁLISE DE ROTINA); UROCULTURA; URÉIA NO SANGUE (NITROGÊNIO UREICO).